TOWARDS EFFICIENT UNCERTAINTY ESTIMATION IN DEEP LEARNING FOR ROBUST ENERGY PREDICTION IN MATERIALS CHEMISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

In material science, recent studies have started to explore the potential of using deep learning to improve property prediction from high-fidelity simulations, e.g, density functional theory (DFT). However, the design spaces are sometimes too large and intractable to sample completely. This results in a critical question that is how to evaluate the confidence and robustness of the prediction. In this paper, we propose an efficient approach to estimate uncertainty in deep learning using a single forward pass and then apply it for robust prediction of the total energy in crystal lattice structures. Our approach is built upon the deep kernel learning (DKL) that originally introduces to leverage the expressiveness of deep neural networks as input with a probabilistic prediction of Gaussian processes (GPs) as output. Existing DKL methods have difficulties in the accuracy of predictive uncertainty, training stability, and scaling to large datasets, which lead to significant barriers in real-world applications. We propose to address these challenges by using an inducing point approximate GP in feature space combined with spectral normalization as a regularization. We finally demonstrate our robust performance on an artificial example and a real-world application from materials chemistry.

1 INTRODUCTION

In the search for advanced materials for various functional applications ranging from energy storage to catalysis, deep learning (DL) has quickly gained traction as a powerful and flexible approach. The adoption of DL for materials design is expected to expand even further with the ongoing growth in the availability of high-throughput density functional theory (DFT) simulation datasets and continued advancements in DL algorithms. This research area has spurred the creation of DL models to predict various material properties, including total energy, bandgap and etc. Unfortunately, in many cases, the design space is very large such that we are unable to sample completely for these DL models. The resulting under-sampling challenge can limit the training data and therefore the predictive capability of the models (Tran et al., 2020). It is useful to have an uncertainty estimation for a DL model so that we can evaluate the robustness and determine when we trust the prediction.

Efficient methods that estimate uncertainty in deep neural networks are critical for real-world applications in science and engineering. A large body of research on this purpose is Bayesian deep learning or Bayesian neural networks (BNNs) (MacKay, 1992; 1995; Neal, 2012) which can be used to interpret the model parameters and robust to over-fitting problems. Even though exact inference in the BNN framework is often tractable, a set of variational approximation methods (Murphy, 2013; Tzikas et al., 2008; Hoffman et al., 2013) are proposed to deal with the scalability challenges (Li & Gal, 2017). More practically, Monte Carlo dropout (Gal & Ghahramani, 2016; 2015) can be seen as a promising alternative way that is easy to evaluate and implement. Another important direction is the deep ensemble methods Lakshminarayanan et al. (2016) that are proposed by combining multiple deep models training from different initializations, and have outperformed the BNN framework that was trained using variational inference (Ovadia et al., 2019). Some very recent methods on deterministic uncertainty quantification (DUQ) (Van Amersfoort et al., 2020; Liu et al., 2020) use a single forward pass and regulates the neural network mapping with a gradient penalty. DUQ methods have scaled well to large image datasets in classification tasks.

In this work, we aim to conduct an efficient uncertainty estimation in deep learning for conducting robust prediction of material properties using DFT simulation datasets. Our approach leverages recent advances in deterministic uncertainty quantification framework and addresses the existing challenges based on the following contributions:

- We combine a residual deep architecture as a feature extractor with an approximate Gaussian process (GP) model to efficiently estimate uncertainty using a single forward pass.
- We propose to introduce inducing point GP with fuzzy c-means clustering that is used to represent the full datasets such that we can reduce the computational complexity.
- We show the robust performance of the approach on total energy prediction in a real-world lattice crystal structure $SrTiO_3$ perovskite oxide from material chemistry.

2 Method

2.1 Deep feature extractor with residual networks

Deep kernel learning (Wilson et al., 2016a) (DKL) is a well-established approach for estimating uncertainty in deep neural networks with a single forward pass. The overall idea of DKL is to first extract the feature by leveraging deep neural networks and then use the feature extractor as an input to a Gaussian process output which offers probabilistic measurement. This is achieved by the use of a kernel that contains a deep feature extractor, which is given by

$$\mathcal{K}_{\tau,\theta}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leftarrow \mathcal{K}^*_\tau(\Omega_\theta(\boldsymbol{x}_i), \Omega_\theta(\boldsymbol{x}_j)) \tag{1}$$

where Ω_θ is a deep neural network parameterized by θ, up to the last linear layer. $\mathcal{K}^*_\tau$ is the base kernel, such as Matérn or RBF kernel that is typically used in GP, and τ is the hyperparameters of the base kernel. In practice, DKL encounters several difficulties, particularly in scaling to large datasets because the exact inference for GP is hampered by the inversion of kernel matrix in Eq. 1, where the time complexity scales cubically with the number of data, N. Thus, DKL is often integrated with a sparse GP approximation (Snelson & Ghahramani, 2005; Quinonero-Candela & Rasmussen, 2005) to mitigate the computational burden.

2.2 Inducing point GP with soft clustering

An alternation method is the use of variational approximation of the exact GP. Titsias (2009) proposed a variational formulation for sparse approximations which jointly infers the kernel parameters τ and inducing inputs by maximizing a lower bound of the log marginal likelihood that is also known as the evidence lower bound, which is defined as

$$L(X) = \log P(\mathbf{X}) - D_{\mathrm{KL}}(Q \parallel P) = -\sum Q(\mathbf{Y}) \log Q(\mathbf{Y}) + \sum Q(\mathbf{Y}) \log P(\mathbf{Y}, \mathbf{X}) \tag{2}$$

where a distribution Q over unobserved variables $\mathbf{Y}$ targets to approximate the true posterior $P(\mathbf{Y}|\mathbf{X})$ given data $\mathbf{X}$. Maximizing the evidence lower bound minimizes the Kullback–Leibler divergence, $D_{\mathrm{KL}}(Q \parallel P)$. Our approach is built upon this idea with a much smaller number of inducing points $m_s \ll N$ to overcome the scalability issue in GP. These inducing points are placed in input feature space to approximate the full dataset and the locations of these points are variational parameters that can be learned by jointly training. Inducing point GP reduces the time complexity of the matrix inversion from $\mathcal{O}(N^3)$ to $\mathcal{O}(m_s^2 N)$.

The number of inducing points may have a sensitive impact on the approximation of the true posterior distribution (Wilson et al., 2016b). van Amersfoort et al. (2021) proposed to place the inducing point in feature space using k-means clustering and demonstrated that only fewer inducing points offer competitive accuracy in image classification tasks. However, the clustering in regression task is often more challenging than classification. Since the dataset may belong to more than one cluster, we propose to use a soft clustering, that is often named as fuzzy c-means (FCM) (Rezaee et al., 1998), which aims to minimize the objective function

$$\mathcal{F} = \underset{C}{\arg\min} \sum_{i=1}^{N} \sum_{j=1}^{m_s} w_{ij}^{m_f} ||\mathbf{x}_i - \boldsymbol{c}_j||^2, \quad w_{ij} \in [0, 1] \tag{3}$$

where $C = \{\boldsymbol{c}_1, ..., \boldsymbol{c}_{m_s}\}$ is the cluster centres, w_{ij} is the partition matrix that quantifies the fuzzy degree to which element $\mathbf{x}_i$ belongs to cluster $\boldsymbol{c}_j$, and $m_f = 2$ is a hyperparameter in the FCM.

2.3 Spectral normalization as a regularization

Another limitation in DKL is the uncertainty estimation is sensitive to changes in the input such that we need to avoid mapping out of distribution data to in distribution feature. This is often called *feature collapse* Van Amersfoort et al. (2020) address this issue by regularising the representation map using a two-side gradient penalty that was first introduced by the improvement of GANs (Gulrajani et al., 2017). van Amersfoort et al. (2021) further improved their previous work by introducing spectral normalization proposed by Miyato et al. (2018) and demonstrated that it is more effective on mitigating feature collapse and faster than the gradient penalty method. In this work, we use spectral normalization as a regularization combined with a residual neural architecture for deep feature extraction. Our proposed algorithm is summarized as Algorithm 1.

Algorithm 1: Efficient uncertainty estimation in deep neural networks

```
Require: training data $\{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N\}$, wide residual neural networks $\Omega_\theta$ with parameters $\theta$, the number of inducing point $m_s$, approximate GP with parameters $\xi$ including inducing point locations $\ell_s$, fuzzy hyperparameter, $m_f$, learning rate $\lambda$.
Initialize inducing points with fuzzy c-means clustering
Draw a random subset of $m_r$ point from the training data $X^{\text{ini}} \subset X$
Compute the fuzzy clustering (soft k-means) on $\Omega_\theta(X^{\text{ini}})$ with $k = m_r$, use the centroids as initial inducing point locations $\ell_s$ in approximate GP.
Train residual neural networks and GP jointly
Implement spectral normalization on residual neural network parameters $\hat{\theta} \leftarrow \theta$
Evaluate forward model to extract the feature space $\psi \leftarrow \Omega_{\hat{\theta}}(\mathbf{x})$
Evaluate approximate GP on feature space with parameter $\xi$, $p(\hat{\mathbf{y}}|\mathbf{x}) \leftarrow \text{GP}_\xi(\psi)$
Define the loss function $\mathcal{L}$ using the negative evidence lower bound, $\mathcal{L} \leftarrow \text{NELB}_\xi(p(\hat{\mathbf{y}}|\mathbf{x}), \mathbf{y})$
Minimize the loss function $\mathcal{L}$ with respect to $\theta$ $\xi$ via $\theta, \xi \leftarrow \theta, \xi + \lambda \cdot \nabla_{\theta,\xi}\mathcal{L}$
```

3 Experiments

3.1 1D sinusoids regression problem

Figure 1 shows the prediction results on a 1D dataset (total 500 data) drawn from sinusoids. Spectral normalization and Matérn kernel in approximate GP is used in this case. We investigate the effect of inducing points size on the prediction and found that fewer points may lead to a bad accuracy but the prediction performance tends to consistent when we increase the number to 50 and 100. There is a trade-off between the accuracy and computational cost so, in practice, we believe we can determine the optimal number of inducing points through several trials without requiring a fine-tune.

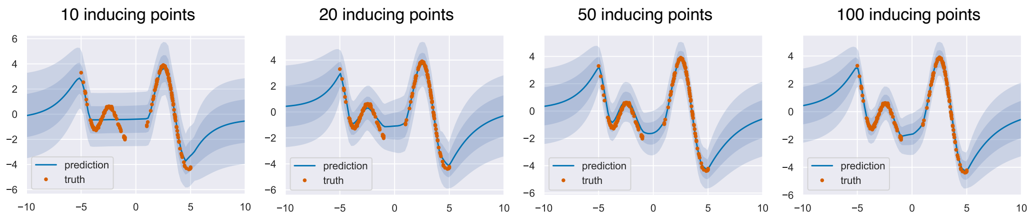

Figure 1: A simple 1D regression task. The true data is marked by red and the the prediction including uncertainty is marked by blue. Different number of inducing points are compared and we can note that the fewer inducing points (e.g., 10 points) may lead to a poor uncertainty estimation and increasing points will improve the prediction including uncertainty as expected.

3.2 Total energy prediction in crystal lattice systems

materials chemistry simulations are performed to simulate these materials under perturbation and obtain their resulting physical properties such as total energy. Here, we tackle the robust prediction of total energy to strain mapping for the case of the SrTiO_3 perovskite oxide, which is otherwise intractable to obtain from materials chemistry. This can be an exceptionally difficult problem due to the complex underlying physics, and the high degree of sensitivity of the total energy to the lattice parameters including the length and angle, requiring very accurate predictions for the generated structures to succeed. This proposed method can be extended to study various crystal lattice structures, including bulk, surface, metal-organic frameworks (MOFs), 2D materials, and cluster structures in materials chemistry, as shown in Fig. 2.

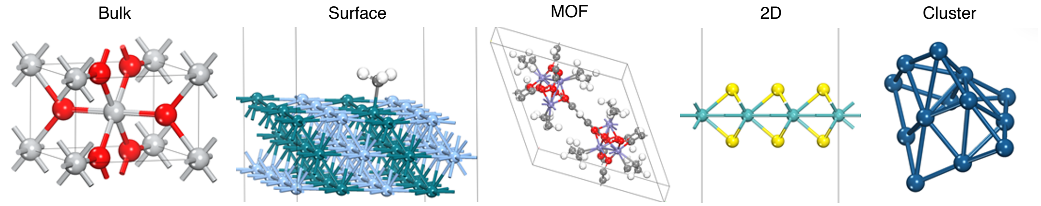

Figure 2: Various crystal lattice structures in materials chemistry.

In this example, the lattice constants and angles a, b, c, α, β, γ were sampled uniformly within a range of 10% deviation from the equilibrium crystal parameters: $a = b = c = 3.914$ Å and $\alpha = \beta = \gamma = 90°$. Within the perturbed ranges (± 10% of equilibrium value) in the lattice constants of the training data, the total energy of the crystals was found to vary between -8 to -37 eV, representing a very wide range in energies. A total of 5000 data (structures) were generated and total energy obtained using DFT which was performed with the Vienna Ab Initio Simulation Package (VASP) package. Here we use spectral normalization, with a Matérn kernel and train 1000 epochs on a Nivida P4000 GPU within 10 minutes.

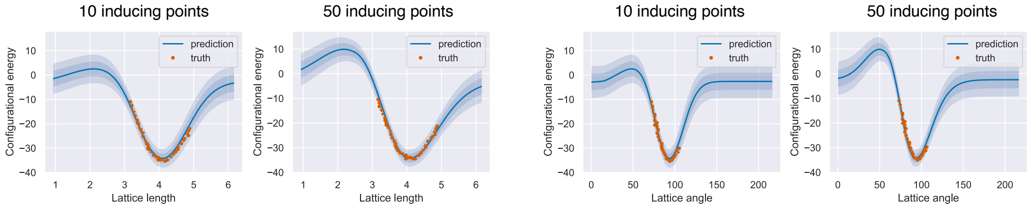

Figure 3: Illustration of the total energy prediction with respect to lattice length and lattice angle parameters given 10 and 50 inducing points.

Fig 3 shows the total energy prediction with respect to the lattice length and angle. Note that our prediction is accurate and robust with low uncertainty near the equilibrium status as discussed above but shows relatively large uncertainty on the input space that is far away from any data seen during training, particularly when lattice length and angle is out of the perturbed range. We also compare the performance caused by the number of inducing points and found that a slight difference exists when there is no data but the prediction is similar within the perturbed range of lattice parameters.

4 Conclusion

We propose an efficient uncertainty estimation in deep learning and apply it for robust prediction of total energy in materials chemistry, specifically crystal lattice systems. Our approach is built upon deep kernel learning (DKL) and addresses the existing challenges by combining spectral normalization and inducing point approximate GP in feature space. The results demonstrate our approach offers a robust prediction for the total energy using DFT simulation data. The future work will aim to compare our approach to other baseline methods, such as Monte Carlo dropout, Bayesian neural networks, deep ensemble methods and deterministic uncertainty estimation methods in terms of the stability, accuracy and efficiency of predictive uncertainty.

REFERENCES

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *arXiv preprint arXiv:1512.05287*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *International conference on machine learning*, pp. 2052–2061. PMLR, 2017.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*, 2020.

David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

David JC MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1):73–80, 1995.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Kevin Murphy. A variational approximation for bayesian networks with discrete and continuous latent variables. *arXiv preprint arXiv:1301.6724*, 2013.

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

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

M Ramze Rezaee, Boudewijn PF Lelieveldt, and Johan HC Reiber. A new cluster validity index for the fuzzy c-mean. *Pattern recognition letters*, 19(3-4):237–246, 1998.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257–1264, 2005.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pp. 567–574. PMLR, 2009.

Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, 2020.

Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pp. 9690–9700. PMLR, 2020.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Improving deterministic uncertainty estimation in deep learning for classification and regression. *arXiv preprint arXiv:2102.11409*, 2021.

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

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Stochastic variational deep kernel learning. *arXiv preprint arXiv:1611.00336*, 2016b.