Efficient Deep Gaussian Process Models for Variable-Sized Inputs

Issam H. Laradji, Mark Schmidt, Vladimir Pavlovic, Minyoung Kim

Motivation

	Tractable	Deep Representation	Fully Bayesian	Sequence Kernels
Gaussian Processes (GP)	True	False	True	True
Deep Gaussian Processes (DGP)	False	True	True	True
Deep Random Feature Expansion (DRF)	True	True	True	False
Deep Neural Networks	True	True	False	False
GP-DRF (Ours)	True	True	True	True

Related Work

Gaussian processes (GP) [Rasmussen and Williams 2006]

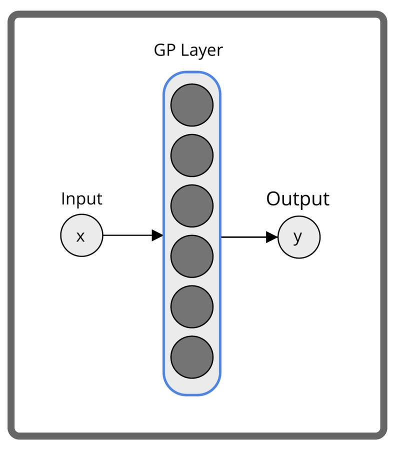
- A single layer of non-parametric latent functions

Deep GP [Damianou and Lawrence 2013]

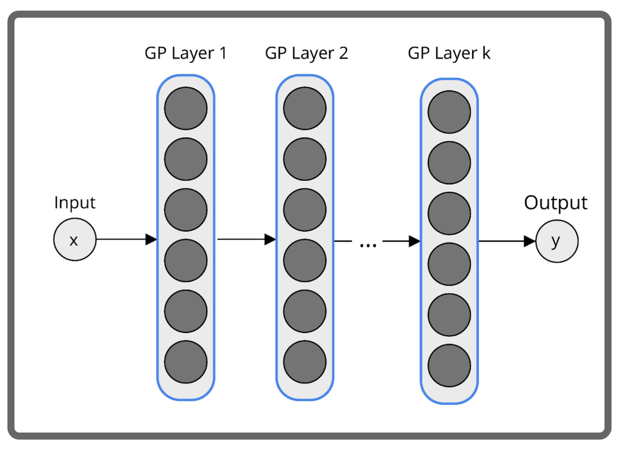
- A sequence of GP layers
- A layer is a latent variable sampled from a GP

Deep random features (DRF) [Cutajer *et al.* 2017]

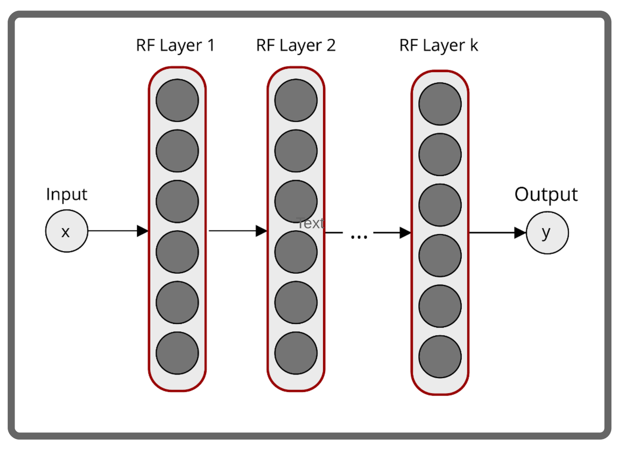
- A sequence of RF layers
- A layer is a linear parametric formulation of a GP



GP



DGP

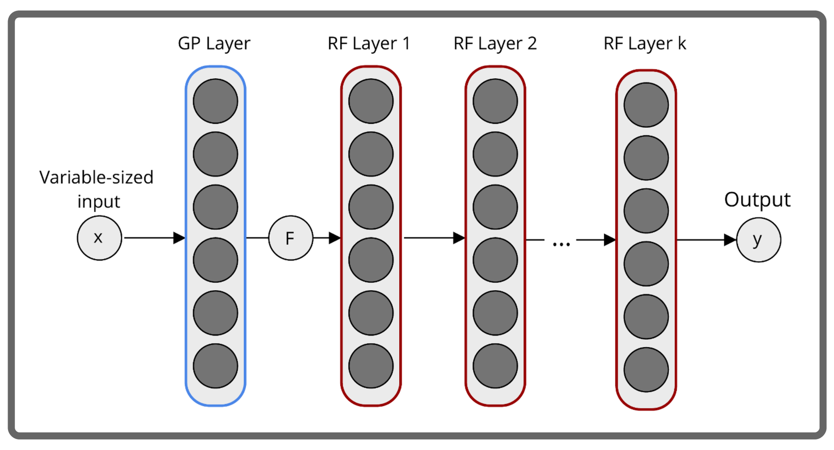


DRF

Method: GP-DRF

We propose **GP-DRF**

- First layer is a GP followed by a set of RF layers,



GP-DRF

Training

- Perform Inference for the following posterior distribution,

$$P(F, W, \Omega | X, Y, \Theta). \qquad (10)$$

- Approximate F, W, Ω using variational density,

$$q(W|\Psi_W) = \prod_{l,i,j} \mathcal{N}(w^l_{i,j}; m^l_{i,j}, (s^l_{i,j})^2) \qquad (12)$$

$$q(\Omega|\Psi_\Omega) = \prod_{l,i,j} \mathcal{N}(\omega^l_{i,j}; \eta^l_{i,j}, (\beta^l_{i,j})^2) \qquad (13)$$

$$q(\overline{F}|\Psi_F) = \prod_{i=1}^{d_0} \mathcal{N}(\overline{F}^j; \mu_j, \Sigma_j), \qquad (14)$$

- Maximize ELBO using SGD,

$$\mathrm{ELBO}(\Psi, \Theta) = \sum_{n=1}^{N} \mathbb{E}_q[\log P(y_n | G(F_n; W, \Omega, \theta_o), \theta_l)] - \mathrm{KL}(q(W, \Omega, \overline{F}) || P(W, \Omega, \overline{F})). \quad (16)$$

Prediction

- Using MCMC, sample $y_*^{(t)} \sim P(y_* | G(F_*^{(s)}; W^{(s)}, \Omega^{(s)}, \theta_o), \theta_l)$
- Estimate the mean and variance, respectively, as,

$$\frac{1}{T}\sum_{t=1}^{T} y_*^{(t)} \quad (:= \overline{y}_*) \quad (27)$$

$$\frac{1}{T-1}\sum_{t=1}^{T} (y_*^{(t)} - \overline{y}_*)^2. \quad (28)$$

Experimental Setup

Two problem setups

- Fixed-sized inputs (FSI)
- Variable-sized inputs (VSI)

Kernel choices

- RBF for FSI
- Double [Kuksa *et al.* 2008] for VSI

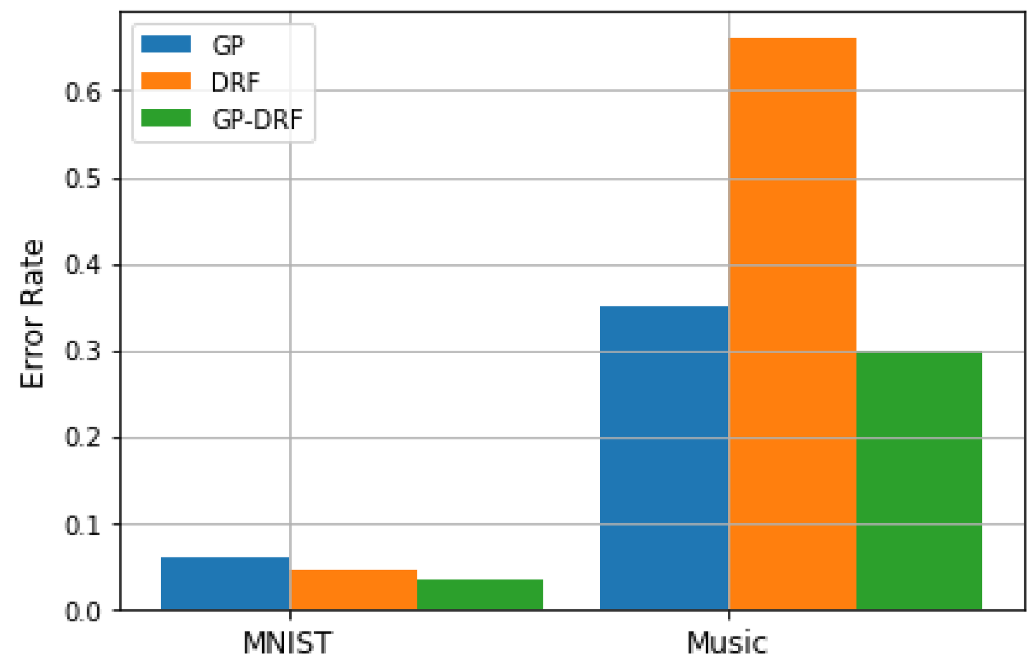
Implementation

- Softmax likelihood
- ADAM with learning rate 1e-5
- 1000 epochs
- 200 Inducing points
- 100 MCMC samples

Results 1: Error Rate

Datasets' description,

	Input	Kernel	# Train	# Test	# Classes
MNIST	Fixed-sized	RBF	60000	10000	10
Music	Variable-sized	Double	900	100	10



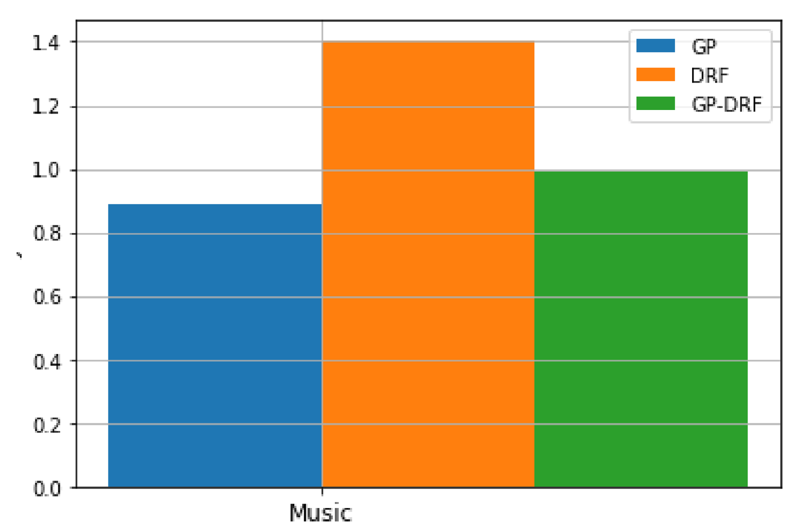
Error rate (Lower is better)

Results 2: Battacharya Distance

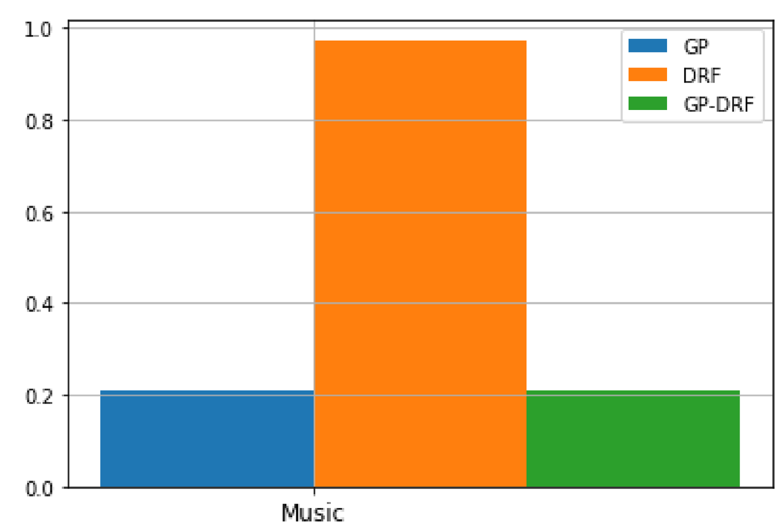
Bhattacharyya Distance [Bhattacharyya 1946]

- Uncertainty analysis to measure the separatibility of classes
- Depends on the mean and standard deviation of the predictions,

$$D(F_*(x), F_+(x)) = \frac{1}{4}\ln\Big(\frac{1}{4}\Big(\frac{\sigma_*}{\sigma_+} + \frac{\sigma_+}{\sigma_*} + 2\Big)\Big) + \frac{1}{4}\Big(\frac{(\mu_* - \mu_+)^2}{\sigma_* + \sigma_+}\Big), \qquad (29)$$



B. Distance for the correctly labeled (Higher is better)



B. Distance for the Incorrectly labeled (Lower is better)

Summary

GP-DRF

- A fast and accurate deep Bayesian model
- GP helps in learning sequence kernels
- DRF for fast approximation of deep Gaussian processes

Future Work

- Use it with more interesting Graph kernels
- Use its uncertainty measure for active learning

Scan me