Continual learning via function-space variational inference

Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, Yarin Gal

Corresponding author: tim.rudner@cs.ox.ac.uk, @timrudner. Paper: https://timrudner.com/cfsvi.

Summary

We formulate continual learning as variational inference over predictive functions, and derive a tractable variational objective for deep neural networks. Our method achieves state-of-the-art performance on benchmark task sequences.

Inference scheme

A variational distribution $q(\theta)$ over the parameters of a neural network induces a variational distribution $q(f)$ over predictive functions.

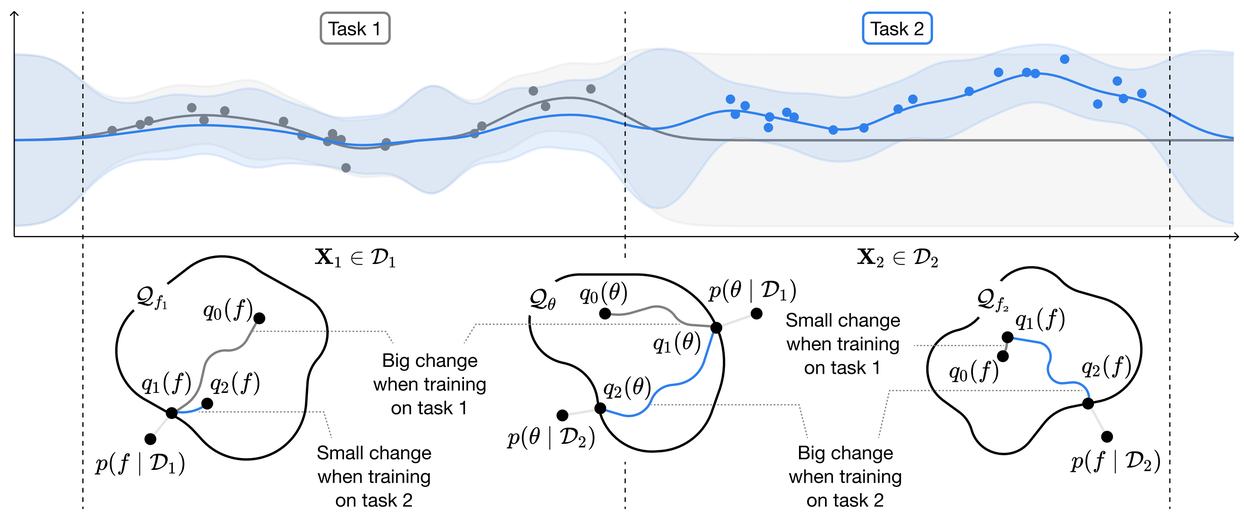
Based on this, we formulate training on the current task t as finding the posterior $q_t(f) \in \mathcal{Q}_f$ that maximizes

$$\mathbb{E}_{q_t(f)}\left[\log p\left(\mathbf{y}_t \mid f\left(\mathbf{X}_t\right)\right)\right] - \mathbb{D}_{\mathrm{KL}}\left(q_t(f) \| q_{t-1}(f)\right)$$

where $(\mathbf{X}_t, \mathbf{y}_t)$ is the data for task t and $q_{t-1}(f)$ is the variational posterior at the end of task $t-1$.

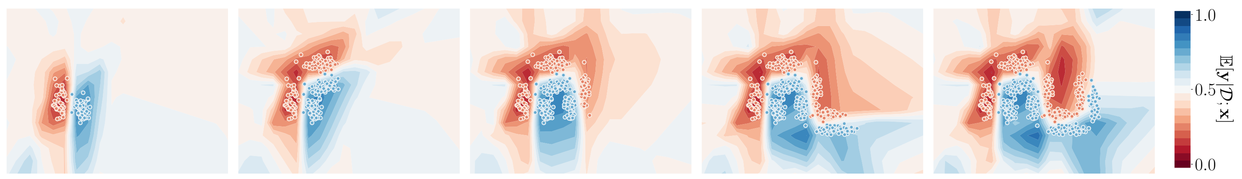
The first term in the objective encourages $q_t(f)$ to fit the data from task t. The second term promotes agreement between $q_t(f)$ and $q_{t-1}(f)$, preventing catastrophic forgetting on previous tasks.

Schematic of training dynamics



On task 1, the variational distribution over functions updates from $q_0(f)$ to $q_1(f)$ to fit dataset $\mathcal{D}_1$. On task 2, the posterior $q_2(f)$ fits dataset $\mathcal{D}_2$ while also matching $q_1(f)$ on a small subset of $\mathcal{D}_1$. The distribution over parameters is free to change significantly since changes in θ are not directly penalized.

Practical demonstration on synthetic data



(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4 (e) Task 5

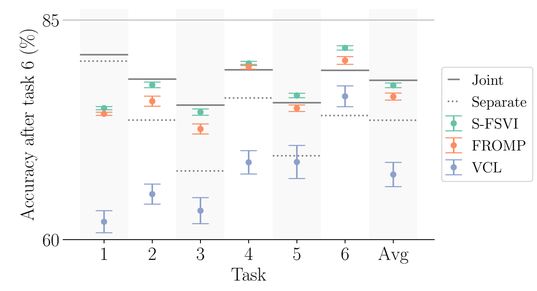
A neural network is tasked with binary classification on data drawn from regions of 2D input space progressively revealed over time. As a result of using sequential function-space variational inference, the neural network infers an accurate decision boundary while maintaining high predictive uncertainty away from the data.

Making inference tractable

In its original form, the variational objective is intractable for neural networks. To resolve this, we

1. Induce a variational distribution over functions by defining a mean-field Gaussian distribution over neural-network parameters
2. Approximate the KL divergence between distributions over functions by linearizing neural networks at parameter means
3. Derive a Monte Carlo estimator for use in stochastic optimization

Performance on split CIFAR



Our method (S-FSVI) outperforms parameter-space variational inference (VCL) and an existing function-space method (FROMP) on split CIFAR, a challenging task sequence.