Approximate Bayesian Inference Team

Mohammad Emtiyaz Khan

Goals and Challenges

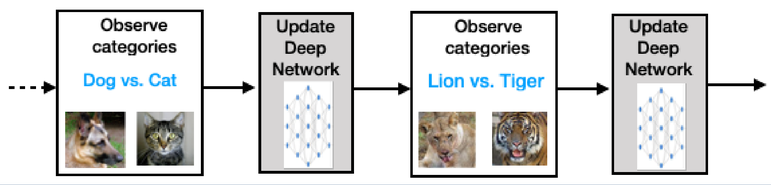
Long-Term Goal: To discover fundamental principles of learning, and design AI systems that continue to learn and improve throughout their life (like humans)

Living beings can learn throughout their life from small chunks of data in a non-stationary world, but deep learning requires a large amount of data from a stationary world. Our current research focuses on reducing this gap.

Main Idea: We use Bayesian principles to enable human-like learning of deep networks.
- New methods for uncertainty estimation in deep networks
- Convert deep networks to Gaussian process to use them as prior for life-long learning.

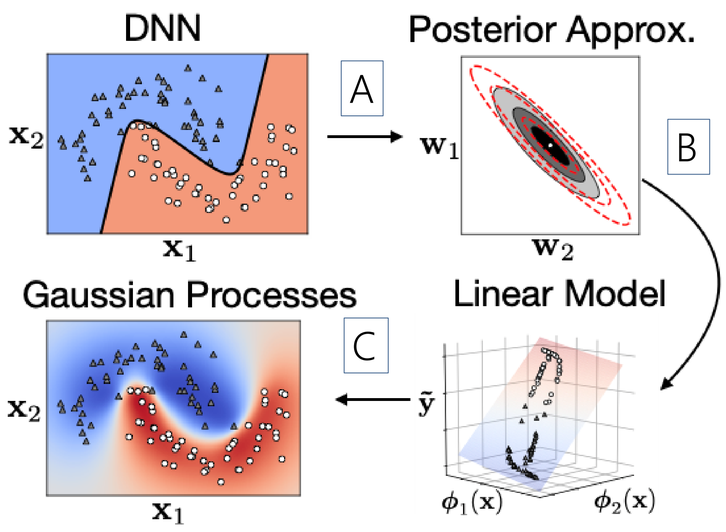
Standard Deep Learning

Deep Continual Learning



From Deep Neural Networks to Gaussian Processes

Using Gaussian posterior approximations, we can convert Neural network to Gaussian processes (GP). This enables us to use deep networks as functional priors.

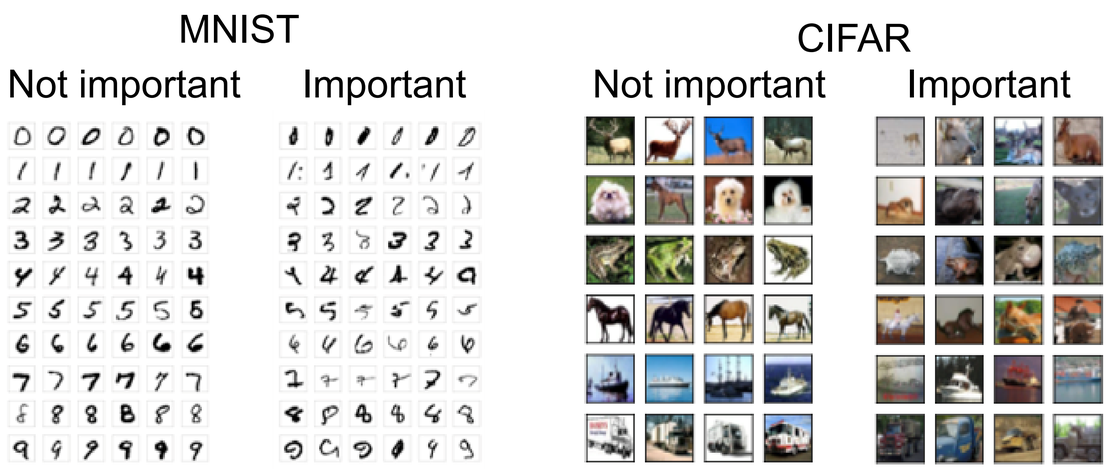


Step A: Find a Gaussian posterior approximation

Step B: Find a linear model whose posterior is equal to the Gaussian approximation

Step C: Find a GP whose predictions are the same as those of the linear model.

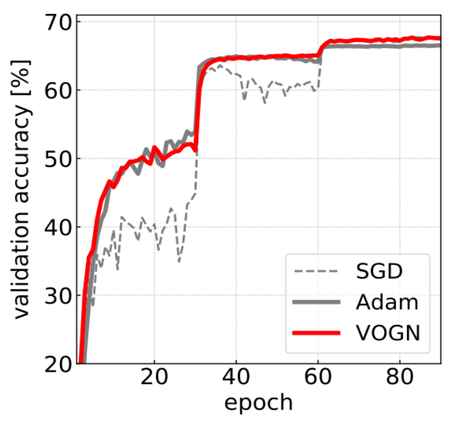
Using this approach, we can identify important data examples.



Approximate Inference Turns Deep Networks into Gaussian Processes. Khan ME, Immer A, Abedi E, Korzepa M (NeurIPS 2019).

Uncertainty Estimation for Large Deep Networks

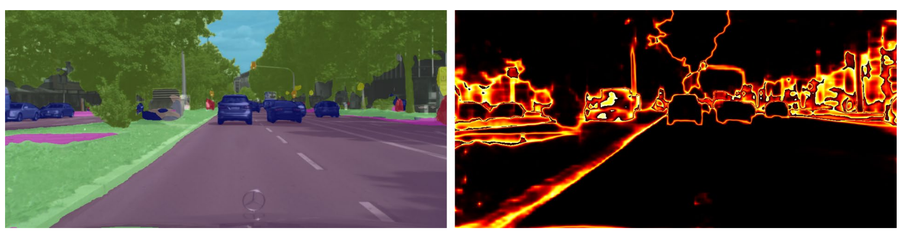
- A new scalable algorithm, called VOGN, to estimate uncertainty for large deep-learning problems.
- For the first time, we can train ResNet-18 on ImageNet with 128 GPUs. We achieve similar performance to Adam/SGD in about the same number of epochs while preserving the benefits of Bayesian principles:
  - predictive probabilities are well-calibrated
  - uncertainties on out-of-distribution data are improved
  - continual-learning performance is boosted



Algorithm Distributed VOGN

repeat
  Sample a minibatch $\mathcal{M}$ of size M.
  Split $\mathcal{M}$ into each GPU (local minibatch $\mathcal{M}_{local}$).
  for each GPU in parallel **do**
    for $k = 1, 2, \ldots, K$ (# MC samples) **do**
      Sample weight $\mathbf{w}^{(k)} \sim q(\mathbf{w})$.
      Compute gradient $\mathbf{g}_i^{(k)} \forall i \in \mathcal{M}_{local}$
      $\hat{\mathbf{g}}_k \leftarrow \frac{1}{M} \sum_{i \in \mathcal{M}_{local}} \mathbf{g}_i^{(k)}$ and $\hat{\mathbf{h}}_k \leftarrow \frac{1}{M} \sum_{i \in \mathcal{M}_{local}} (\mathbf{g}_i^{(k)})^2$.
    end for
    $\hat{\mathbf{g}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \hat{\mathbf{g}}_k$ and $\hat{\mathbf{h}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \hat{\mathbf{h}}_k$.
  end for
  AllReduce (aggregate) $\hat{\mathbf{g}}, \hat{\mathbf{h}}$ among all GPUs.
  $\mathbf{m} \leftarrow \beta_1 \mathbf{m} + (\hat{\mathbf{g}} + \tilde{\delta}\mu)$ and $\mathbf{s} \leftarrow (1 - \tau\beta_2)\mathbf{s} + \beta_2 \hat{\mathbf{h}}$.
  $\mu \leftarrow \mu - \alpha \mathbf{m} / (\mathbf{s} + \tilde{\delta})$.
until stopping criterion is met =0

Bayesian Uncertainty Estimation in Image Segmentation



Segmentation (left) and its predictive entropy (right) on Cityscapes dataset

Practical Deep Learning with Bayesian Principles. Osawa K, Swaroop S, Jain A, Eschenhagen R, Turner RE, Yokota R, Khan ME (NeurIPS 2019).

A New Bayesian Learning Rule for Mixture of Exponential Family

- Bayesian Learning rule is a general learning rule from which many learning-algorithms can be derived
  - Deep learning (SGD and Adam)
  - Least-squares, Kalman filters etc.
  - And many more..
- We extended the application of rule to mixture of exponential family, enabling us to derive, e.g., an ensemble Newton method.

Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations. Lin W, Khan ME, Schmidt M (ICML 2019).

Inference Networks for Gaussian-Process Models

- This work uses neural networks to approximate the posterior distribution of Gaussian process models.
- The main idea is to use a ”functional” mirror descent algorithm.
- The updates can be approximated using a neural networks as the posterior distribution.

Scalable Training of Inference Networks for Gaussian-Process Models. Shi J, Khan ME, Zhu J (ICML 2019).

A Generalization Bound for Online Variational Inference

- This work derives new bounds for methods that perform online variational inference.
- We considers a variety of such algorithms and showed that the generalization error is bounded, even though each step of the algorithm is approximate.
- Best paper award at ACML 2019.

A Generalization Bound For Online Variational Inference. Chérief-Abdellatif, BE, Alquier, P and Khan, ME (ACML 2019).