Bayesian Deep Learning

Maurizio Filippone

EURECOM, Sophia Antipolis, France

January 10th, 2019
1 Motivation

2 Probabilistic Deep Nets
   Scalable Inference
   Connections with (Deep) Gaussian Processes

3 Some Results

4 Conclusions
Motivation
Quantification of Uncertainty with Expensive Models

- Climate modeling

Quantification of Uncertainty with No Models

- Classification and progression of neurodegenerative diseases

Healthy?

Needs treatment?

Filippone et al., AoAS, 2012 – Lorenzi, Filippone et al., NeuroImage, 2017 – Lorenzi and Filippone, ICML, 2018
A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate
A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A model might not even be available

- Replace it with a flexible statistical model
A Unified Framework

A model might be expensive to simulate/inaccurate

• Emulate model/discrepancy using a surrogate

A model might not even be available

• Replace it with a flexible statistical model

Probabilistic Deep Models for Accurate Modeling and Quantification of Uncertainty
Probabilistic Deep Nets
- Take these two examples

- We are interested in estimating a function $f(x)$ from data
- Most problems in Machine Learning can be cast this way!
Deep Neural Networks

- Implement a composition of parametric functions

\[ f(x) = f^{(L)} \left( f^{(L-1)} \left( \ldots f^{(1)}(x) \ldots \right) \right) \]

with

\[ f^{(l)}(h) = g \left( h^\top W^{(l)} \right) \]
Back-propagation – Probabilistic Interpretation Loss

- **Inputs**: $X = \{x_1, \ldots, x_N\}$
- **Labels**: $Y = \{y_1, \ldots, y_N\}$
- **Weights**: $W = \{W^{(1)}, \ldots, W^{(L)}\}$

Quadratic Loss

\[
p(Y|X, W) \propto \exp(-\text{Loss})
\]

- Back-propagation minimizes a loss function
- \ldots equivalent as optimizing likelihood $p(Y|X, W)$
Bayesian Inference

- **Inputs**: \( X = \{x_1, \ldots, x_N\} \)
- **Labels**: \( Y = \{y_1, \ldots, y_N\} \)
- **Weights**: \( W = \{W^{(1)}, \ldots, W^{(L)}\} \)

\[
p(W|Y, X) = \frac{p(Y|X, W)p(W)}{\int p(Y|X, W)p(W)dW}
\]
• Regression example
• Classification example
Bayesian inference is intractable due to this integral:

\[
\log [p(Y|X)] = \log \left[ \int p(Y|X, W)p(W)dW \right]
\]

Lower bound for \( \log [p(Y|X)] \):

\[
E_{q(W)}[\log [p(Y|X, W)] - KL[q(W) || p(W)]
\]

where \( q(W) \) approximates \( p(W|Y, X) \).

Kullback-Leibler divergence (KL) is the "distance" between \( q \) and \( p \).

Optimize the lower bound with respect to the parameters of \( q(W) \).
• Bayesian inference is intractable due to this integral

$$\log [p(Y \mid X)] = \log \left[ \int p(Y \mid X, W)p(W)dW \right]$$

• Lower bound for $$\log [p(Y \mid X)]$$

$$E_{q(W)} (\log [p(Y \mid X, W)]) - KL [q(W) \parallel p(W)] ,$$

where $$q(W)$$ approximates $$p(W \mid Y, X)$$.

• Kullback-Leibler divergence $$KL$$ – “distance” between $$q$$ and $$p$$
Stochastic Variational Inference

- Bayesian inference is intractable due to this integral

$$\log [p(Y|X)] = \log \left[ \int p(Y|X, W)p(W)dW \right]$$

- Lower bound for $\log [p(Y|X)]$

$$E_{q(W)} \left( \log [p(Y|X, W)] \right) - KL [q(W)\|p(W)]$$

where $q(W)$ approximates $p(W|Y, X)$.

- Kullback-Leibler divergence $KL$ – “distance” between $q$ and $p$

Optimize the lower bound wrt the parameters of $q(W)$
• Assume that the likelihood factorizes

\[ p(Y|X, W) = \prod_k p(y_k|x_k, W) \]
Stochastic Variational Inference

- Assume that the likelihood factorizes

\[ p(Y|X, W) = \prod_k p(y_k|x_k, W) \]

- Doubly stochastic unbiased estimate of the expectation term
  - Mini-batch

\[ E_{q(W)} \left( \log [p(Y|X, W)] \right) \approx \frac{n}{m} \sum_{k \in \mathcal{I}_m} E_{q(W)} \left( \log [p(y_k|x_k, W)] \right) \]

- Monte Carlo

\[ E_{q(W)} \left( \log [p(y_k|x_k, W)] \right) \approx \frac{1}{N_{MC}} \sum_{r=1}^{N_{MC}} \log[p(y_k|x_k, \tilde{W}_r)] \]

with \( \tilde{W}_r \sim q(W) \).
Assume a factorized Gaussian approximate posterior:

\[ q(W) = \prod_{ijl} q(W_{ij}^{(l)}) = \prod_{ijl} \mathcal{N}(\mu_{ij}^{(l)}, (\sigma^2)_{ij}^{(l)}) \]  

(1)
Stochastic Variational Inference

- Assume a factorized Gaussian approximate posterior:

\[ q(W) = \prod_{ijl} q(W_{ij}^{(l)}) = \prod_{ijl} \mathcal{N}(\mu_{ij}^{(l)}, (\sigma^2)^{ij}^{(l)}) \]  

(1)

- Reparameterization trick

\[ (\tilde{W}_r^{(l)})_{ij} = \sigma_{ij}^{(l)} \varepsilon_{rij} + \mu_{ij}^{(l)} , \]

with \( \varepsilon_{rij} \sim \mathcal{N}(0, 1) \)

- Optimization wrt \( \mu_{ij}^{(l)}, (\sigma^2)^{ij}^{(l)} \) with automatic differentiation

Kingma and Welling, *ICLR*, 2014
Stochastic Gradient Optimization

\[ E \left\{ \nabla_{\text{par}_q} \text{LowerBound} \right\} = \nabla_{\text{par}_q} \text{LowerBound} \]

---

Robbins and Monro, AoMS, 1951
\[
\text{par}_q' = \text{par}_q + \frac{\alpha_t}{2} \nabla_{\text{par}_q} \text{(LowerBound)} \quad \alpha_t \to 0
\]
Approximating distribution $q(W)$ can have the following forms:

- Fully factorized

\[
\begin{pmatrix}
\begin{array}{c}
-5 \\
0 \\
5 \\
\end{array}
\end{pmatrix}
\]
Approximating distribution $q(W)$ can have the following forms:

- Fully factorized

- Full covariance $\Sigma = LL^T$
Approximating distribution $q(W)$ can have the following forms:

- Fully factorized
- Full covariance $\Sigma = LL^T$
- Normalizing Flows, Real NVPs, Stein VI - change of measure determined by $\det(\text{Jacobian})$

Initialization of SVI matters

- Initialization can be an issue
- We proposed a novel way to initialize SVI well

Rossi, Michiardi and Filippone, arXiv, 2018
Is There an Easier Way?

- Dropout is Variational Inference with Bernoulli-like $q(W)$
- At training time, apply dropout
  
  Iteration 1     Iteration 2     Iteration 3 ... 

- At test time, “sample” networks with different dropout masks

• Take $W^{(i)} \sim \mathcal{N}(0, \alpha_i I)$
• Central Limit Theorem implies that $F$ is Gaussian

• $F$ has zero-mean
• $\text{cov}(F) = E_p(W^{(0)}, W^{(1)}) \left[ \Phi(X W^{(0)}) W^{(1)} W^{(1)\top} \Phi(X W^{(0)})^\top \right]$
• Take $W^{(i)} \sim \mathcal{N}(0, \alpha_i I)$
• Central Limit Theorem implies that $F$ is Gaussian

• $F$ has zero-mean
• $\text{cov}(F) = \alpha_1 E_p(W^{(0)}) [\Phi(XW^{(0)}) \Phi(XW^{(0)})^\top]$ 
• Some choices of $\Phi$ lead to analytic expression of known kernels (RBF, Matérn, arc-cosine, Brownian motion, ...)

Random Feature Expansions for DGPs - Bochner’s theorem

- Continuous shift-invariant covariance function

\[ k(x_i - x_j | \theta) = \sigma^2 \int p(\omega | \theta) \exp \left( i (x_i - x_j)^T \omega \right) d\omega \]
Random Feature Expansions for DGPs - Bochner’s theorem

- Continuous shift-invariant covariance function

\[ k(x_i - x_j | \theta) = \sigma^2 \int p(\omega | \theta) \exp \left( \iota (x_i - x_j)^\top \omega \right) d\omega \]

- Monte Carlo estimate

\[ k(x_i - x_j | \theta) \approx \frac{\sigma^2}{N_{RF}} \sum_{r=1}^{N_{RF}} z(x_i | \tilde{\omega}_r)^\top z(x_j | \tilde{\omega}_r) \]

with

\[ \tilde{\omega}_r \sim p(\omega | \theta) \]

\[ z(x | \omega) = [\cos(x^\top \omega), \sin(x^\top \omega)]^\top \]

Random Feature Expansions for DGPs

- Define

\[ \Phi(l) = \sqrt{\frac{\sigma^2}{N_{RF}^{(l)}}} \left[ \cos \left( F^{(l)} \Omega^{(l)} \right), \sin \left( F^{(l)} \Omega^{(l)} \right) \right] \]

and

\[ F^{(l+1)} = \Phi(l) W^{(l)} \]

- We are stacking Bayesian linear models with

\[ p \left( W^{(l)} \right) = \mathcal{N} (\mathbf{0}, I) \]
Random Feature Expansions for DGPs

• Define

\[
\phi^{(l)} = \sqrt{\frac{\sigma^2}{N_{RF}^{(l)}}} \left[ \cos \left( F^{(l)} \Omega^{(l)} \right), \sin \left( F^{(l)} \Omega^{(l)} \right) \right]
\]

and

\[
F^{(l+1)} = \phi^{(l)} W^{(l)}
\]

• We are stacking Bayesian linear models with

\[
p \left( W^{(l)} \right) = \mathcal{N} \left( 0, I \right)
\]

• Expansion of arc-cosine kernel yields ReLU activations!

Cutajar, Bonilla, Michiardi, Filippone, ICML, 2017
Random Feature Expansions make Deep GPs become DNNs

Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017
Some Results
Airline dataset

\((n = 5M+, \, d = 8)\)
Convolutional nets are widely used...
...but they are known to be overconfident!
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams
● Reliability diagrams
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams - Under-confident predictions

- We can extract the Expected Calibration Error (ECE) score
- The BRIER score is another measure of calibration
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams - Overconfident predictions

Reliability diagrams of modern Deep CNNs look like this!
Bayesian treatment of filters fixes it!
Calibration as a Measure of Quantification of Uncertainty

- Reliability diagrams - Overconfident predictions

Reliability diagrams of modern Deep CNNs look like this! Bayesian treatment of filters fixes it!
Bayesian CNNs are calibrated

- Inferring parameters of convolutional filter recovers calibration
- Example with Monte Carlo Dropout

RELIABILITY DIAGRAM FOR MCD

Tran et al., *AISTATS*, 2019
• Bayesian CNNs are calibrated and achieve better performance than post calibrated CNNs
Knowing When the Model Doesn’t Know

- Training on MNIST and test on not-MNIST

![Graph showing MCD and I-BLM Initialization](image)

Conclusions
Conclusions

- Inference for Deep Nets is hard
  - Scalable stochastic-based approximate inference but...
  - ... it is difficult to assess the impact approximations on quantification of uncertainty
Conclusions

- Inference for Deep Nets is hard
  - Scalable stochastic-based approximate inference but...
  - ... it is difficult to assess the impact approximations on quantification of uncertainty
- The connection between Deep Nets and Deep Gaussian processes can have implications on
  - Understanding Deep Learning
  - Deriving sensible priors for Deep Learning
  - Improving inference borrowing algebraic/computational tricks from kernel literature
Conclusions

- Inference for Deep Nets is hard
  - Scalable stochastic-based approximate inference but...
  - ... it is difficult to assess the impact approximations on quantification of uncertainty
- The connection between Deep Nets and Deep Gaussian processes can have implications on
  - Understanding Deep Learning
  - Deriving sensible priors for Deep Learning
  - Improving inference borrowing algebraic/computational tricks from kernel literature
- Cool stuff
  - New hardware
  - Bayesian compression
We are hiring PhDs, Post-docs and Assistant Professors
Thank you!