Bayesian Deep Learning

Presented by Roberto Halpin Gregorio

Primary Papers

Wilson, A. G. and Izmailov, P. NeurIPS 2020.

Bayesian Deep Learning and a Probabilistic Perspective of Generalization.

Wilson, A. G. 2020.

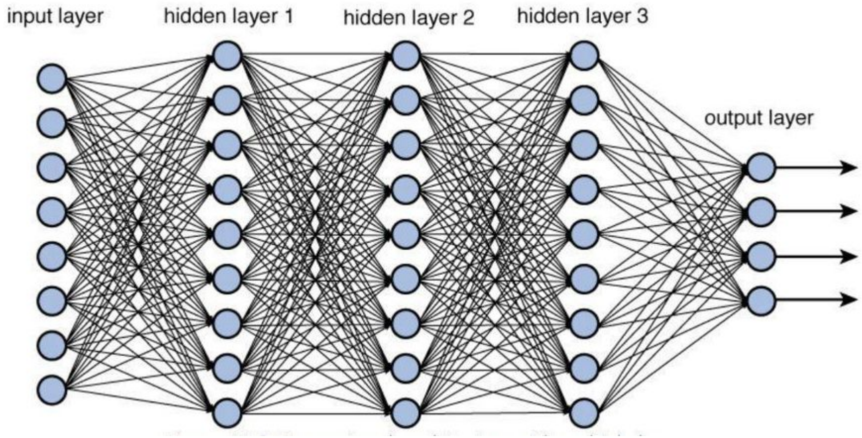
The Case for Bayesian Deep Learning. 2020.

Deep Learning

- Models based on the composition of many parameterized function modules trained from examples using gradient-based optimization.

- Very powerful and popular, but mysterious modern machine learning method.

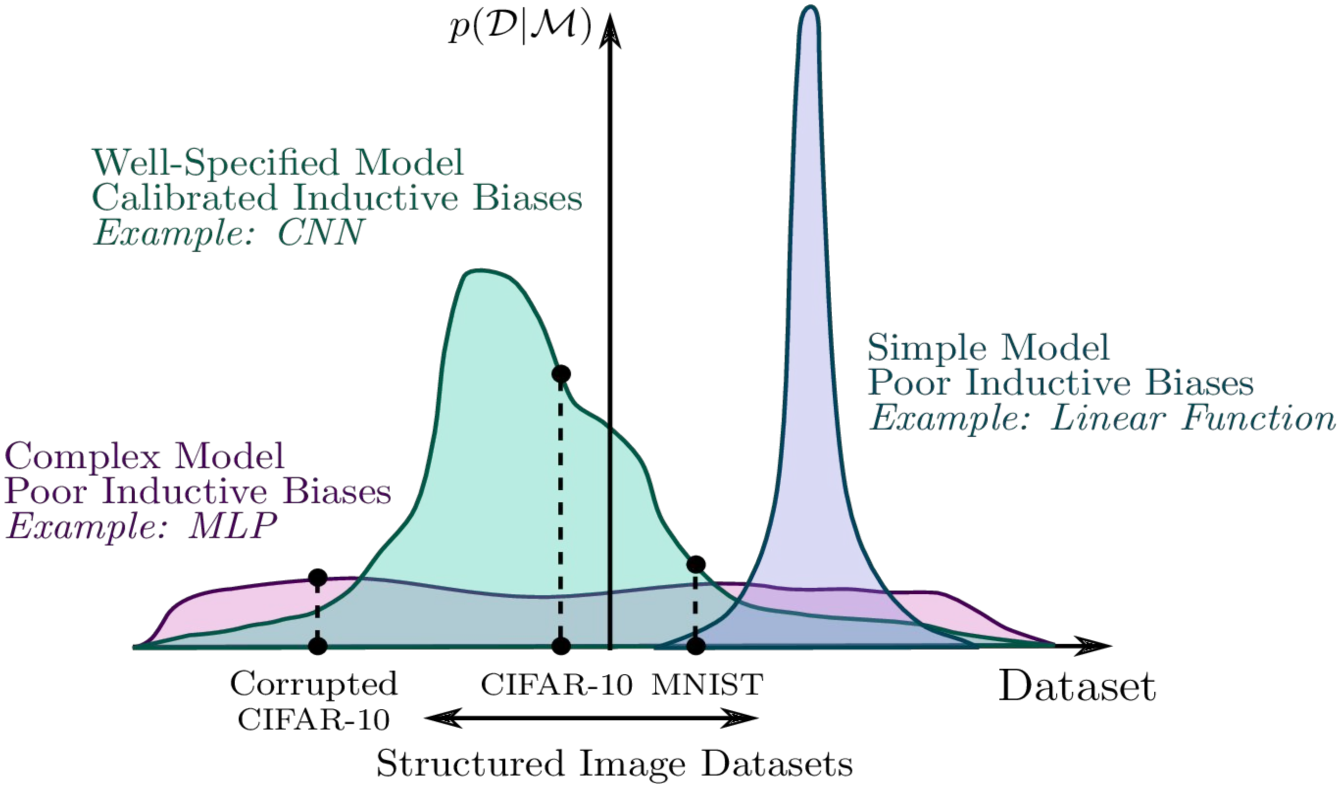
- Heavily used in Computer Vision, Natural Language Processing, and many other fields.



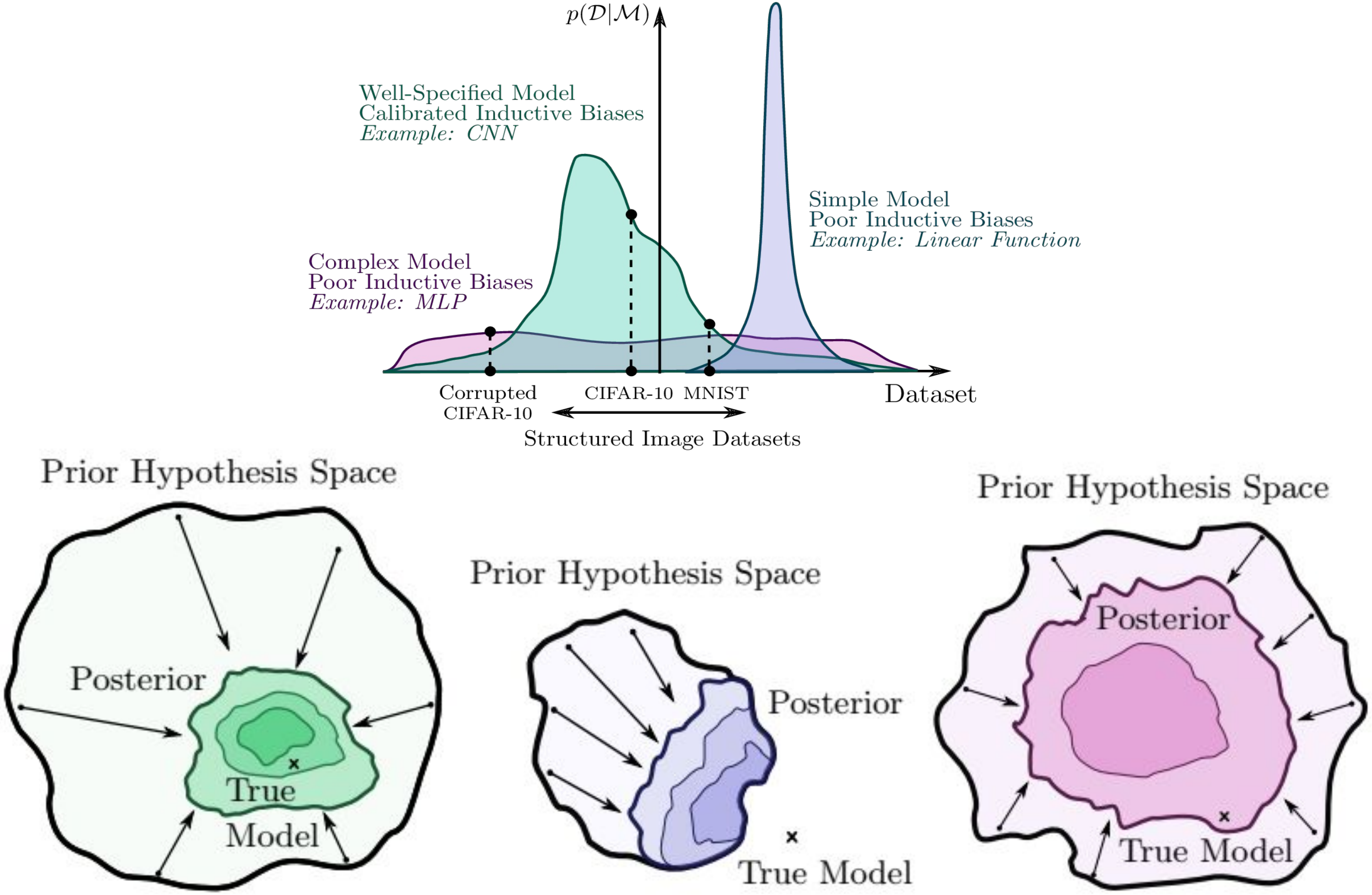
Generalization

- "The evidence, or marginal likelihood, $p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\mathcal{M}, w)p(w)dw$, is the probability we would generate a dataset if we were to randomly sample from the prior over functions $p(f(x))$ induced by a prior over parameters $p(w)$."
- Inductive biases are "the relative the relative prior probabilities of different datasets — the distribution of support given by $p(\mathcal{D}|\mathcal{M})$."
- "The support is the range of datasets for which $p(\mathcal{D}|\mathcal{M}) > 0$."
- Deep Learning models have a large support and thus fit many datasets.

Wilson, A. G. and Izmailov, P. 2020.



$p(\mathcal{D}|\mathcal{M})$
Well-Specified Model
Calibrated Inductive Biases
Example: CNN
Simple Model
Poor Inductive Biases
Example: Linear Function
Complex Model
Poor Inductive Biases
Example: MLP
Corrupted
CIFAR-10
CIFAR-10
MNIST
Dataset
Structured Image Datasets



CIFAR-10 Dataset

Bayesian Approach (marginalization)

We want to compute the Bayesian model average (BMA):

$$p(y|x, \mathcal{D}) = \int p(y|x, w)p(w|\mathcal{D})dw$$

y - outputs (e.g., regression values, class labels, . . .)
x - inputs (e.g. spatial locations, images, . . .)
w - weights (or parameters) of the model
$\mathcal{D}$ - data

Instead of using a single setting of parameters, we use all possible parameter settings weighed by their posterior probabilities.

Classical vs. Bayesian Approach

BMA: $p(y|x, \mathcal{D}) = \int p(y|x, w)p(w|\mathcal{D})dw$

- Classical training can be seen as using $p(w|\mathcal{D}) \approx \delta(w = \hat{w}) \quad \hat{w} = \text{argmax}_w p(w|\mathcal{D})$
  - Leads to standard predictive distribution $p(y|x; \hat{w})$

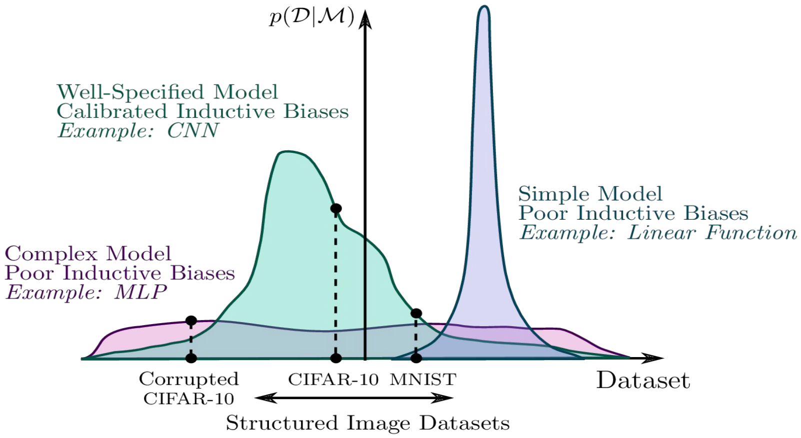
- If the actual posterior is not unimodal with a sharp peak, then the delta function is not a reasonable approximation.

- Bayesian Deep Learning: using the Bayesian model average (BMA for deep learning models.

The case for Bayesian Deep Learning

- Neural networks tend to be underspecified by the data.
  - Many more parameters than data.
  - Leads to diffuse likelihoods $p(\mathcal{D}|w)$, which do not favor any one set of parameters.

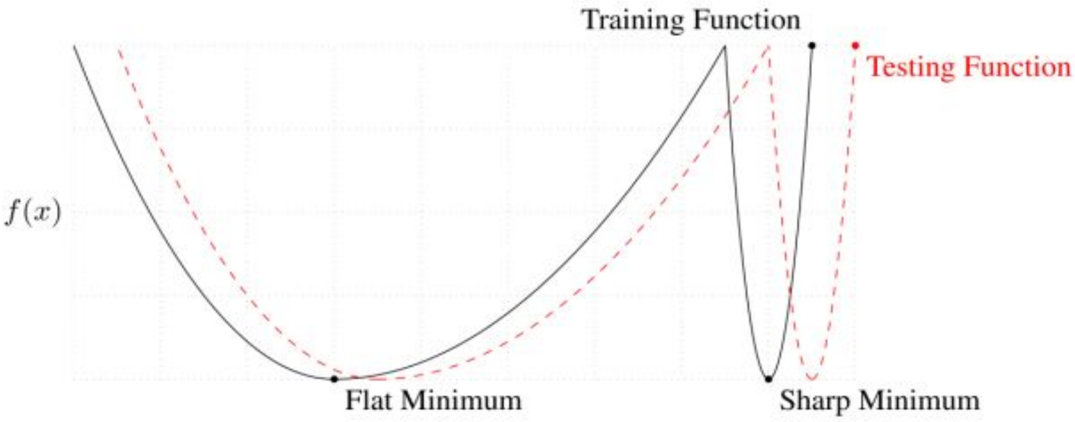
- Many different high performing models corresponding to different settings of parameters.*



*Garipov et al., 2018; Izmailov et al., 2019

The case for Bayesian Deep Learning

- Solutions in flat regions of the posterior correspond to better generalization [1].
- These flat solutions take up much more volume in high dimensions [2].



[1] Garipov et al., 2018; Izmailov et al., 2018 [2] Huang et al., 2019 Diagram: Keskar et. al, 2017

The case for Bayesian Deep Learning

- Uncertainty representation
  - Examine the spread of the predictive distribution, $p(y|x;w)$.

- Improved accuracy
  - Averaging the predictions of multiple, accurate models that disagree in some cases should lead to improved accuracy.
  - Empirically shown in Deep Ensembles and Subspace inference.

- Explainability due to the probabilistic underpinnings
  - Bayesian model average is a statement in probability.

Computing (approximate inference)

- BMA: $p(y|x,\mathcal{D}) = \int p(y|x,w)p(w|\mathcal{D})dw$
  - Very non-convex posterior landscape and a very high dimensional parameter space.
  - Not analytic (for most models).
- Solution: Simple Monte Carlo approximation

$$p(y|x,\mathcal{D}) \approx \frac{1}{J}\sum_{j} p(y|x,w_j), \quad w_j \sim q(w|\mathcal{D})$$

w_j are samples from an approximate posterior $q(w|\mathcal{D})$.

Approximate Posterior $q(w|\mathcal{D})$

- Deterministic Methods
  - Approximate $p(w|\mathcal{D})$ with $q(w|\mathcal{D},\theta)$, usually Gaussian.
  - Examples:
    - Laplace, Expectation Propagation, Variational, Standard Training.

- MCMC
  - Create a Markov chain of approximate samples from $p(w|\mathcal{D})$.
  - Examples:
    - Metropolis-Hastings, Hamiltonian Monte Carlo (HMC), Stochastic gradient HMC, Stochastic gradient Langevin dynamics.

Downsides of Bayesian Deep Learning

- Computational cost

- Computational intractability
  - No exact solution to the Bayesian model average.

- Many design decisions
  - Aproximate Inference method.
  - More hyperparameters.

References

- Primary Sources

  - Wilson, A. G. and Izmailov, P. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. 2020.
  - Wilson, A. G. The Case for Bayesian Deep Learning. 2020.
  - Wilson, A. G. Bayesian Deep Learning and a Probabilistic Perspective of Model Construction. ICML 2020 Tutorial.

- Supplementary

  - Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. 2018.
  - Huang, W. R., Emam, Z., Goldblum, M., Fowl, L., Terry, J. K., Huang, F., and Goldstein, T. Understanding generalization through visualizations. 2019.
  - Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. Subspace inference for Bayesian deep learning. 2019.

  - Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. Subspace inference for Bayesian deep learning. 2019.

  - Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: