Fast Uncertainty Estimates and Bayesian Model Averaging of DNNs

Wesley Maddox1 Timur Garipov2,3 Pavel Izmailov1 Dmitry Vetrov4,5 Andrew Gordon Wilson1

1Cornell University, 2Samsung AI Center in Moscow, 3Lomonosov Moscow State University,
4Higher School of Economics, 5Samsung-HSE Laboratory.

Motivation and Contribution

We want to capture information about the uncertainty of deep neural network (DNN) predictions.

We extend Stochastic Weight Averaging (SWA) [1] by forming a Gaussian distribution around the SWA mean.

SWA-Gaussian (SWAG) produces reliable uncertainty estimates, while maintaining accuracy in Bayesian model averaging.

Methods

Stochastic weight averaging [1] uses the average of the weights of SGD to compute predictions for DNNs: \( b_{\text{SWA}} = \frac{1}{K} \sum_{i=1}^{K} \theta_i \).

For convex (and other nice problems), Polyak-Ruppert averaging is asymptotically normal: \( \bar{\theta} \sim N(\theta_{\text{true}}, H^{-1}SH^{-1}) \), motivating the use of a Gaussian distribution [2].

SGD iterates with a constant learning rate are also thought to behave in an approximately Gaussian manner [4].

- **SWAG**: \( \theta \sim N(\theta_{\text{SWA}}, XX^T) \), \( X_i = (\theta_i - \theta_{\text{SWA}}) \).
- **SWAG-Diagonal**: \( \theta \sim N(\theta_{\text{SWA}}, \sum_{i=1}^{K} \theta_i^2 - \theta_{\text{SWA}}^2) \).

Use as an approximate posterior distribution over \( \theta \)

Other Gaussian posterior approximations:

- **Laplace**: \( N(\theta_{\text{MAP}}, \sigma H^{-1}) \) (\( H^{-1} \) is very expensive...)
- **Variational Bayes**: \( N(\mu, \Sigma) \) (which \( \mu, \Sigma ? \))

![Model Calibration](image)

**Bayesian Model Averaging**

Training only requires memory overhead (for storage).

Test time is just \( K \) forwards passes (+ cheap sampling).

Predictions are made using Bayesian model averaging:

\[
p(y^*|y) = \mathbb{E}_{p(\theta|y)}[p(y^*|\theta)] = \frac{1}{K} \sum_{i=1}^{K} p(y^*|\theta_i), \quad \theta_i \sim q_{\text{SWAG}}(\theta|Y).
\]

![Bayesian Model Averaging](image)

Out of Distribution Uncertainty

We trained VGG16 on 5 classes from CIFAR10, and then tested on all 10 classes.

Entropy, \( \sum_{i=1}^{10} p(y = i) \log p(y = i) \), should be higher if the model is unsure.

![Out of Distribution Uncertainty](image)

Conclusions

- Principled method for approximate Bayesian inference that scales well for DNNs.
- SGD posterior appears Gaussian, as theory predicts [2,4]; can also interpret SWA as a posterior mean.

Future Work

- Expand theoretical motivation, like in [3].
- Comparisons with other methods for approximate Bayesian inference – Laplace, Variational Bayes, MC Dropout, etc...
- Other problems: penalized regression [3], adversarial robustness, ImageNet, image segmentation

Code

https://github.com/wjmaddox/swa_uncertainties

References


Acknowledgements

Ruqi Zhang, Jacob Gardner. WSM is supported by an NSF Graduate Research Fellowship.