2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)

FlexGibbs: Reconfigurable Parallel Gibbs Sampling Accelerator for Structured Graphs

Glenn G. Ko*, Yuji Chai†, Rob A. Rutenbar‡, David Brooks*, Gu-Yeon Wei*
*Harvard University, Cambridge, USA
†University of Illinois at Urbana-Champaign, Urbana, USA
‡University of Pittsburgh, Pittsburgh, USA
gko@seas.harvard.edu, ychai6@illinois.edu, rutenbar@pitt.edu, {dbrooks, guyeon}@eecs.harvard.edu

The recent surge of machine learning deployment has motivated a new wave of research in the hardware community for accelerating machine learning, especially for deep learning. The latest variants of GPUs are modified to better support deep learning workloads and major cloud service providers are employing ASIC or FPGA accelerators on the cloud. While these hardware innovations paired with software frameworks that allows quick description of deep neural networks have resulted in systems with orders of magnitude improvements, the same cannot be said for a class of machine learning algorithms such as Bayesian modeling and inference.

Bayesian models and inference are useful as they can perform semi-supervised or unsupervised learning with little or no training data. They are also capable of representing and manipulating uncertainties within the model, which is often important for domains with inherent uncertainties. However, Bayesian models are not yet well supported by existing software frameworks for deep learning and the inference often involve kernels that are not necessarily efficient on GPUs. There has been growing interest in improving the describability of the models through probabilistic programming languages and accelerating the inference by studying new architectures.

One of the most commonly used Bayesian inference algorithm called Gibbs sampling is a Markov chain Monte Carlo (MCMC) method that sequentially samples a variable given all the other variables. Its simplicity and versatility promoted a wide usage for finding the distribution of a probabilistic models in various domains. We studied Gibbs sampling inference on a 2D-grid structured graphical model, Markov random field (MRF), which is widely used for various perceptual tasks, such as stereo matching, image segmentation and source separation. Fig. 1 shows the sampling process on a node in MRF.

FlexGibbs is a flexible accelerator based on parameterized architecture for parallel Gibbs sampling inference on MRFs. Based on the user-configured parameters, the number of labels and MRF dimension supported, optimal Gibbs sampler array is generated for a given FPGA fabric. It also includes a scheduler to perform chromatic scheduling on the nodes onto the array of samplers using 2-coloring of nodes to create two conditionally independent sets of nodes that can be sampled in parallel and converge to the correct stationary distribution [1].

This work was supported by the Applications Driving Architectures (ADA) Research Center, a JUMP Center co-sponsored by SRC and DARPA.

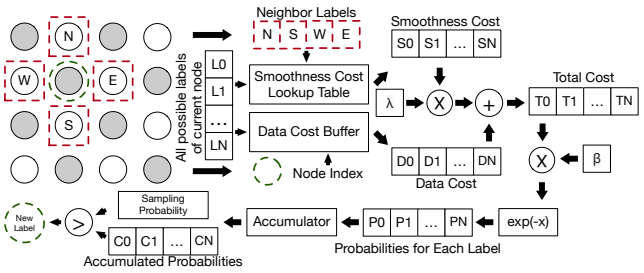

Fig. 1. The process of sampling a new label for a node in an MRF is shown. The data cost from the sampled node is combined with smoothness cost found using neighbors' labels, to find the conditional distribution for each label. A uniform random number is scaled to find $SamplingProbability$, a threshold used to sample a new label by comparing the cumulative sums of the conditional distribution.

Using FlexGibbs, we created a 24x513 binary label MRF Gibbs sampling inference accelerator to be used to accelerate sound source separation task [2]. The larger size of 24x513 resulted in improved separation quality over the prior state-of-the-art hardware for real-time sound source separation using MRF of 8x513. [3]. A 256 array of Gibbs samplers is generated with the chromatic scheduler used to schedule conditionally independent nodes. It effectively samples all nodes of first color of 2-coloring in parallel as the MRF is processed row by row. Once the end is reached, the nodes of the second color are sampled, completing a iteration of Gibbs sampling.

FlexGibbs configured on the FPGA logic of Xilinx Zynq UltraScale+ ZCU102-ES2 board achieved Gibbs sampling inference speedup of 1048x and energy reduction of 99.85% over ARM Cortex-A53 for 50 iterations of Gibbs sampling.

References

[1] J. Gonzalez, Y. Low, A. Gretton, and C. Guestrin, "Parallel gibbs sampling: From colored fields to thin junction trees," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 324–332.

[2] M. Kim, P. Smaragdis, G. G. Ko, and R. A. Rutenbar, "Stereophonic spectrogram segmentation using markov random fields," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, Sept 2012, pp. 1–6.

[3] G. G. Ko and R. A. Rutenbar, "Real-time and low-power streaming source separation using markov random field," *J. Emerg. Technol. Comput. Syst.*, vol. 14, no. 2, pp. 17:1–17:22, May 2018. [Online]. Available: http://doi.acm.org/10.1145/3183351

2576-2621/19/$31.00 ©2019 IEEE
DOI 10.1109/FCCM.2019.00075