SUPPLEMENTARY INFORMATION

Zeinab Abboud[1] Herve Lombaert[1] and Samuel Kadoury[1 2]

[1] Polytechnique Montreal, Montreal, QC, Canada
[2] CHUM Hospital Research Center, Montreal, QC, Canada

Table 1: Segmentation performance on LIDC-IDRI and ISIC datasets and classification results on ChestMNIST, comparing different benchmarks with the proposed partial Bayesian approach with varying r_{bayes}. The fully Bayesian model was trained over 200 epochs to achieve comparative results.

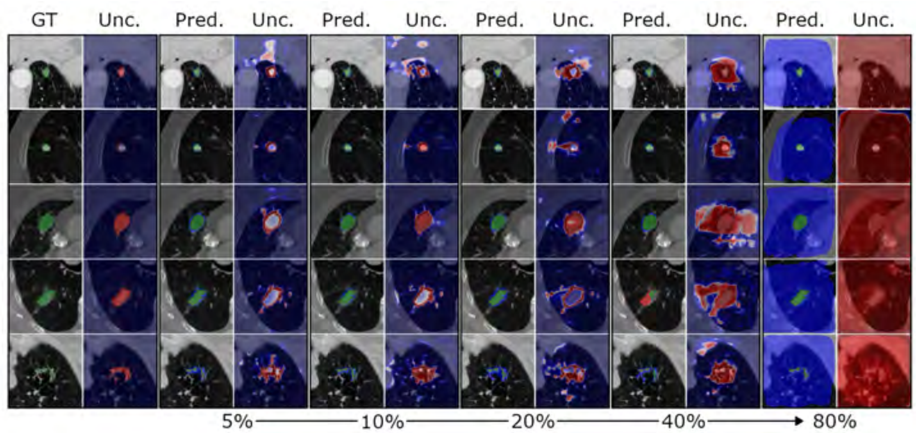
Dataset	**Metric**	**Deterministic**	**Ensemble**	**Partial 1%**	**Partial 5%**	**Partial 10%**
Chest MNIST	**Accuracy** ↑	0.899	0.936	0.934	0.931	0.932
	AUC ↑	0.696	0.726	0.682	0.677	0.684
	Brier Score ↓	0.098	0.053	0.064	0.067	0.072
	Entropy↓	0.493	1.162	0.367	0.404	0.591
	ECE ↓	0.09 ± 0.06	0.05 ± 0.03	0.05 ± 0.04	0.06 ± 0.05	0.05 ± 0.04
LIDC-IDRI	**Dice** ↑	0.71 ± 0.01	0.687 ± 0.002	0.80 ± 0.01	0.75 ± 0.02	0.75 ± 0.02
	Brier Score ↓	0.0746 ± 0.0002	0.0745 ± 0.0001	0.004 ± 0.001	0.005 ± 0.001	0.006 ± 0.001
	Entropy ↓	0.0068 ± 0.0004	0.0076 ± 0.0002	0.010 ± 0.001	0.016 ± 0.003	0.024 ± 0.005
	ECE ↓	0.0048 ± 0.0006	0.0045 ± 0.0005	0.004 ± 0.001	0.005 ± 0.001	0.007 ± 0.003
ISIC	**IoU** ↑	0.801	0.803	0.783	0.762	0.733
	Brier Score ↓	0.110	0.109	0.068	0.076	0.087
	Entropy ↓	0.151	0.144	0.154	0.197	0.238
	ECE ↓	0.027	0.038	0.027	0.040	0.061
	FLOPs ↓	1×	5×	3.36×	3.45×	3.57×

Dataset	**Metric**	**Partial 20%**	**Partial 40%**	**Partial 80%**	**Bayesian**
Chest MNIST	**Accuracy** ↑	0.925	0.783	0.519	0.723
	AUC ↑	0.688	0.729	0.553	0.674
	Brier Score ↓	0.083	0.182	0.267	0.215
	Entropy↓	1.303	4.06	4.787	3.794
	ECE ↓	0.06 ± 0.04	0.27 ± 0.08	0.44 ± 0.05	0.3 ± 0.1
LIDC-IDRI	**Dice** ↑	0.75 ± 0.02	0.68 ± 0.07	0.40 ± 0.06	0.67 ± 0.09
	Brier Score ↓	0.013 ± 0.003	0.02 ± 0.01	0.23 ± 0.07	0.11 ± 0.05
	Entropy ↓	0.05 ± 0.01	0.07 ± 0.04	0.51 ± 0.06	0.3 ± 0.1
	ECE ↓	0.02 ± 0.01	0.02 ± 0.02	0.33 ± 0.09	0.12 ± 0.06
ISIC	**IoU** ↑	0.677	0.653	0.644	0.675
	Brier Score ↓	0.110	0.118	0.121	0.111
	Entropy ↓	0.300	0.318	0.331	0.273
	ECE ↓	0.126	0.123	0.129	0.081
	FLOPs ↓	3.8×	4.27×	5.2×	$> 15\times$

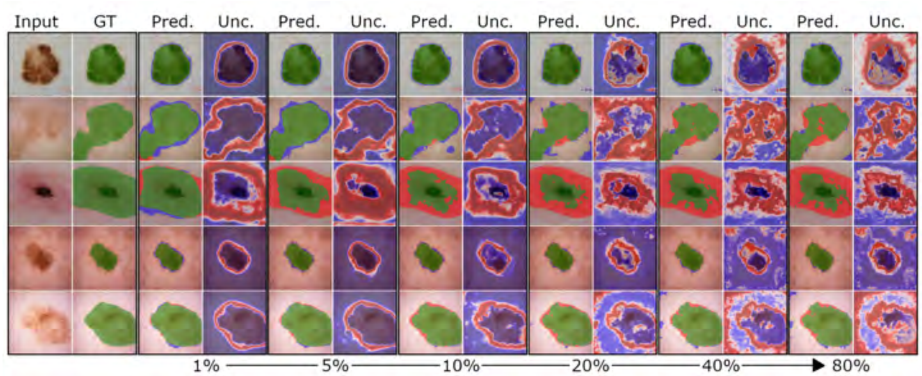
1 FLOPs Count

Given a point-estimate model with θ parameters, its associated training FLOPs count is $3f_d$. The relative ensemble parameter and FLOPs count are $M\theta$ and $3Mf_d$, respectively. The parameter count for a fully Bayesian variational inference model is 2θ, with FLOPs $2 \times (N+2)f_d$. For a Partial Bayesian model with $(1 + r_{\text{bayes}})\theta$ parameters, the FLOPs count is $3f_d \times \text{epochs}_{\text{pretrain}} + (1 + r_{\text{bayes}})(N + 2)f_d$.

2 Qualitative Segmentation Examples



(a) LIDC-IDRI



(b) ISIC

Fig. 1: LIDC-IDRI segmentation predictions for partial Bayesian models with varying percentage of Bayesian parameters 1%, 5%, 10%, 20%, 40%, and 80%, with input image and ground truth segmentation (GT), and input and ground truth uncertainty based on the 4 expert raters. Predictions mask overlays show the true positive (green), false positive (blue), and false negative (red). At the same time, the uncertainty map is the entropy of the output probability, showing regions of high uncertainty (red) and low uncertainty (blue). The figure highlights that fewer selected Bayesian parameters result in better performance and uncertainty representation with a given computational budget.