In Depth Semantic Scene Completion [Bayesian CNN]

David Gillsjö, Kalle Åström

Centre for Mathematical Sciences, Lund University, Sweden

Introduction

Semantic Scene Completion (SSC) is a challenging task in which both visible and occluded surfaces are labeled semantically in 3D. In Figure 1 we see an illustration of the problem where a UAV would benefit from knowing what to expect in occluded areas.

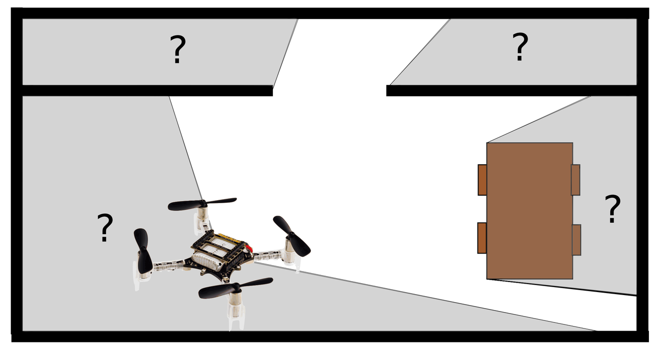


Figure 1: An UAV has some occluded areas in its surrounding and would like to have an idea about what to expect.

Our contributions include:

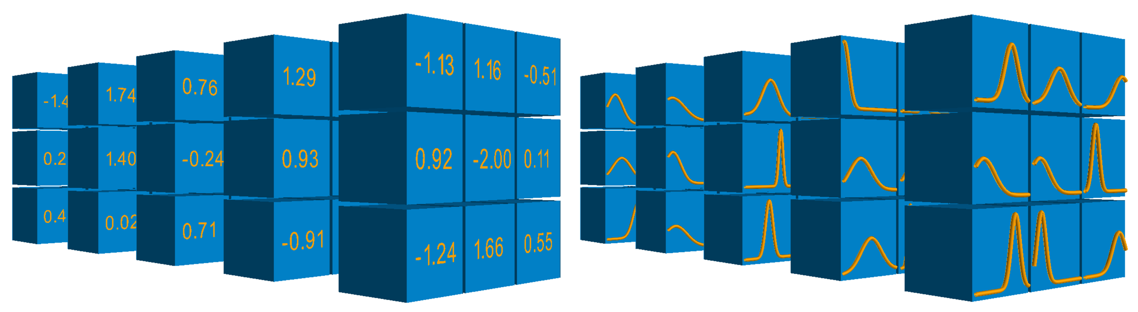
- An open source system for BSSC using Variational Inference released on `https://github.com/DavidGillsjo/bssc-net`.
- An extended SSC task on the SUNCG dataset with more occluded space.
- Experiments showing that the Bayesian approach is more robust to unseen data in the SSC task.
- Parameter studies on both MNIST and SUNCG.

Bayes by backprop

This method introduced by [1] is based on Variational Inference. Each weight in the network is sampled from a normal distribution, as illustrated in Figure 2.
We estimate the posterior $\mathrm{P}(\mathbf{w}|\mathcal{D})$ using a simpler model $\mathrm{q}(\mathbf{w}|\theta)$ with learnable parameters θ, which minimizes the approximate Kullback-Leibler (KL) divergence to the true posterior.

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \frac{\beta}{n} \underbrace{\left[\log \mathrm{q}(\mathbf{w}^{(i)}|\theta) - \log \mathrm{P}(\mathbf{w}^{(i)})\right]}_{\text{Complexity}} - \underbrace{\log \mathrm{P}(\mathcal{D}|\mathbf{w}^{(i)})}_{\text{Likelihood}}.$$

where $\mathbf{w}^{(i)}$ is a sample from the variational posterior $\mathrm{q}(\mathbf{w}^{(i)}|\theta)$. The scale factor $\frac{\beta}{n}$ with β as design parameter is introduced to tune the regularization.

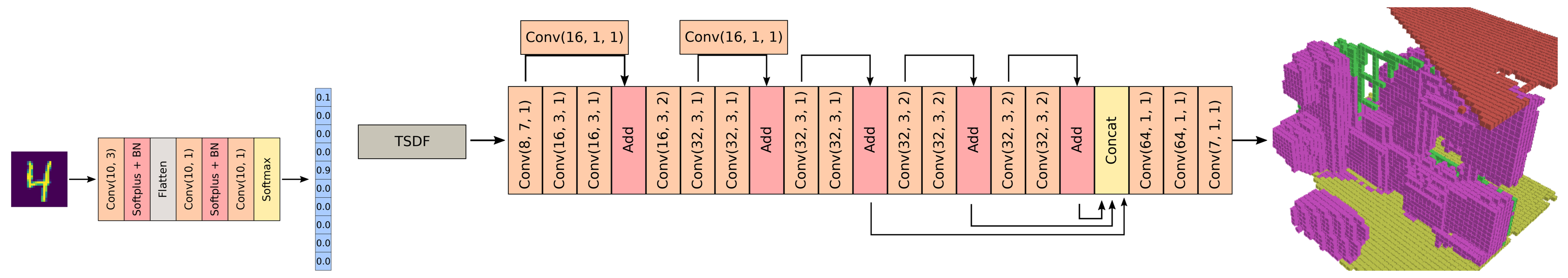


(a) Standard (b) Bayesian

Figure 2: In 2a we see a filter bank from a standard 2D CNN, each weight is a scalar. In 2b we see a filter bank in a Bayesian Variational Inference 2D CNN, here each weight represented as a distribution which is sampled from at inference time.

Model

We have explored two network architectures. The first network architecture is inspired by the original SUNCG article [2]. We call it **SSC-Net**. The second architecture is a **UNet**. We chose softplus as activation functions instead of relu to have more active weights in the network [3]. The architecture is displayed in Figure 3.



(a) MNIST (b) SSC-Net

Figure 3: Architecture of SSC-Net used for MNIST and SUNCG experiments. Conv(d, k, l) stands for a 3D convolution filter stack of depth d and kernel size k and dilation l. Batch normalization and softplus activation is performed after every Conv layer. Softmax in the final layer.

Prediction, Uncertainty & Metrics

An unbiased estimation of the expectation is given [3] by

$$\mathbb{E}_{\mathrm{q}(\mathbf{w}|\theta)}\left[\mathrm{P}(\hat{\mathbf{y}}|\hat{\mathbf{x}}, \mathbf{w})\right] = \int \mathrm{q}(\mathbf{w}|\theta)\mathbf{p}_t \, d\mathbf{w} \approx \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}_t,$$

where $\mathbf{p}_t := \mathrm{P}(\hat{\mathbf{y}}|\hat{\mathbf{x}}, \mathbf{w}^{(t)})$ is the softmax output from forward pass t.
For uncertainty we use Predictive Entropy,

$$\mathrm{H} = -\sum_{t=1}^{T} \mathbf{p}_t \log \mathbf{p}_t.$$

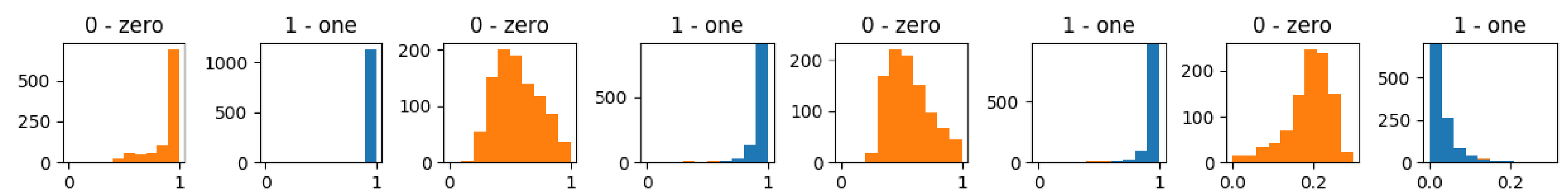
For metrics we use mean Average Precision (mAP), Intersection over Union (IoU) for performance. For separation metric we use the Bhattacharyya coefficient (BC)

$$\mathrm{BC}(\mathbf{p}, \mathbf{q}) = \frac{1}{N}\sum_{i=1}^{N} \sqrt{p_i q_i},$$

where N is the number of categories, q_i and p_i are the number of TP and FN. Lower score indicates better separation.

MNIST Experiment

In Figure 4 we see output distributions from MNIST test set for digits 0 and 1 when 0 is removed from the training data. The Bayesian Score is more better calibrated and the Entropy is higher for 0.



(a) Deterministic (b) Deterministic Regularized (c) Bayesian Mean Score (d) Bayesian Entropy

Figure 4: Here we see true (blue) and false (orange) predictions for 0 and 1.

SUNCG Experiment

SUNCG [2] is a large dataset with manually created and labeled synthetical indoor scenes. We've used a subset of 2000 training and 1000 testing scenes for the experiments. In Figure 5 we show a parameter study on β.Figure 6 shows example output. We also conducted an experiment when category *bed* was removed from training, the result is presented in Table 1.

Table 1: BC, mAP and mIoU for different network architectures when the *bed* class is removed from training. S=Score, E=Entropy. We observe that Bayesian SSC-Net has the best score in all metrics.

CNN	mIoU	mAP: S	mAP: E	BC: S	BC: E
SSC-Net ω=0	0.19	0.2		0.31	
SSC-Net ω=0.01	0.14	0.23		0.29	
B-SSC-Net	**0.21**	**0.26**	**0.19**	**0.27**	**0.28**

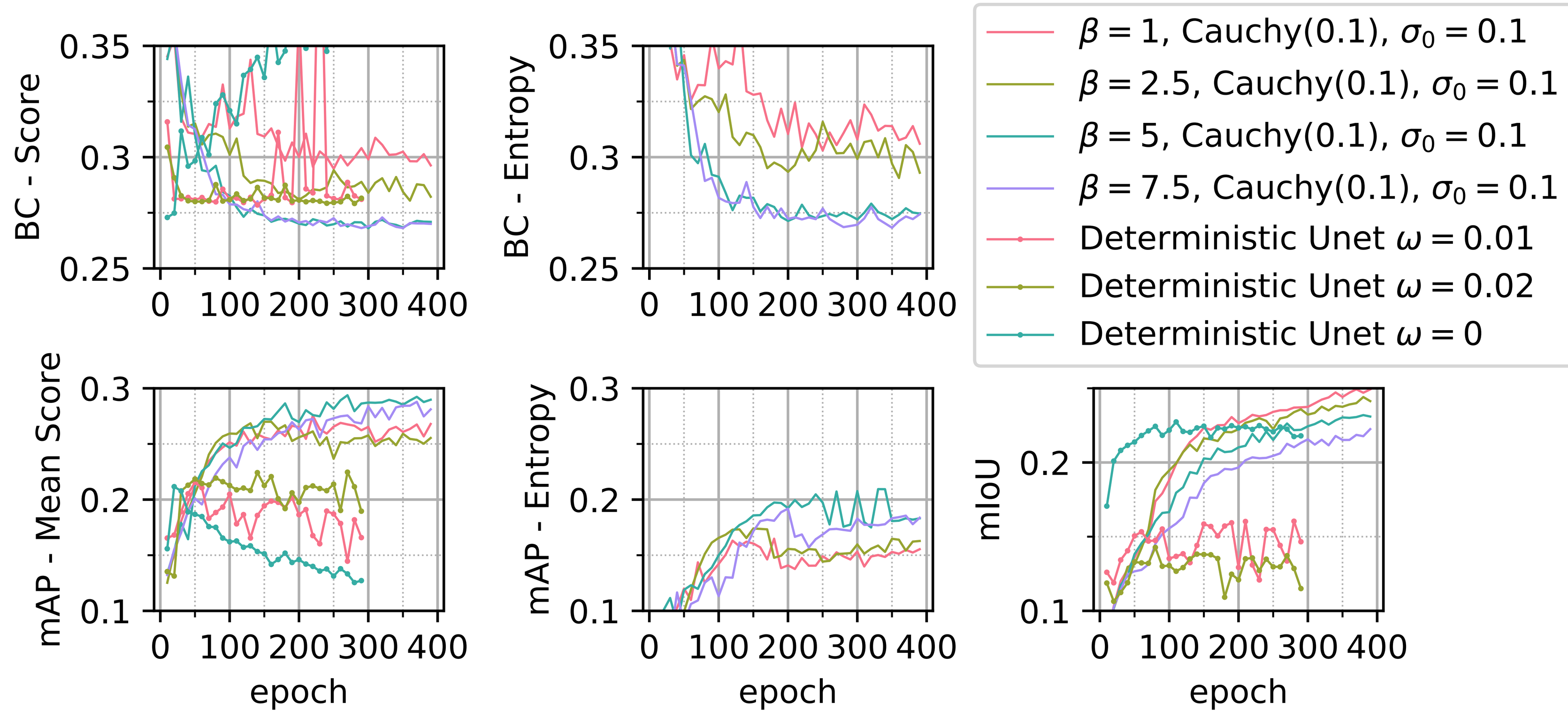


Figure 5: BC, mAP and mIoU for the Bayesian UNet with different weights β and ω for the SUNCG mini dataset. We observe that $\beta = 5$ is better in all metrics but mIoU, where $\beta = 1$ is best.

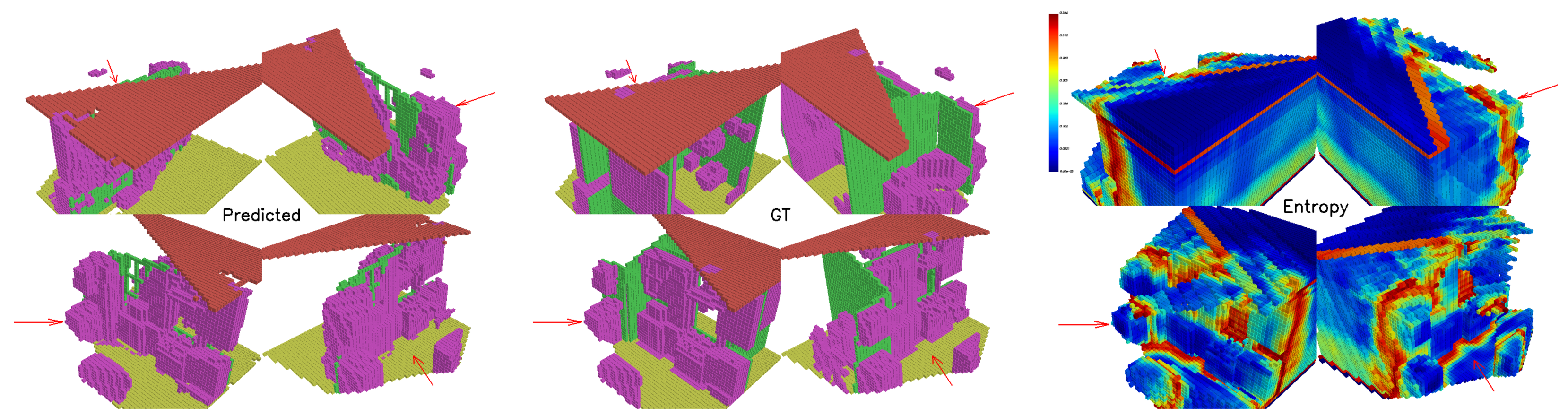


Figure 6: Example from the SUNCG test set. From the left we have predicted, true labels and entropy.

Selected References

See paper for more details and more experiments on MNIST & SUNCG!

[1] C. Blundell, J. Cornebise, K. Kavukcuoglu and D. Wierstra, 'Weight uncertainty in neural networks', *arXiv preprint arXiv:1505.05424*, 2015.

[2] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva and T. Funkhouser, 'Semantic scene completion from a single depth image', *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] K. Shridhar, F. Laumann and M. Liwicki, *A comprehensive guide to bayesian convolutional neural network with variational inference*, 2019. arXiv: 1901.02731 [cs.LG].