Supplementary Material for "Neural Rate Estimator and Unsupervised Learning for Efficient Distributed Image Analytics in Split-DNN models"

Training details

Hyper-parameter	Classification	Segmentation
Optimizer	SGD	ADAM
Learning Rate	0.5	1E-3
Scheduler	Step	Step
Scheduler Rate	15	10(S), 1(US)
Epochs	30 (S), 20(US)	30(S), 1(US)
Batch Size	96	8

We present detailed results here on the outcomes of the hyperparameter search procedure outlined in Section 2.2 of the paper. As described, for each task we perform search the 4-dimensional hyperparameter space (C_r, S, L, Q) over the ranges provided in Table I using random search. We perform this search for each task (classification and segmentation) for both supervised and unsupervised scenarios at different split points. For each 4-tuple, we train the bottleneck layers and measure the resultant task metric (accuracy or mIOU) and rate (bpp, or bits-per-pixel). We then generate a scatter plot of the task metric against the bpp, from which we extract the Pareto frontier. The points on this frontier represent the optimal trade-off between task performance and compression level. The scatter plots, Pareto frontiers, along with the hyperparameter value for each point on the Pareto frontier are provided below.

Architectural detail of hyper prior network

It is an autoencoder network consisting of two conv layers at the hyperprior encoder and two deconv layers in the decoder, viz. Conv1 $(M \times 3 \times 3/1\downarrow)$, Conv2 $(8 \times 1 \times 1/1\downarrow)$, and DeConv1 $(M \times 1 \times 1/1\uparrow)$, DeConv2 $(M \times 3 \times 3/1\uparrow)$ respectively, where M is the number of channels in the output of the bottleneck encoder.

The convention followed to denote a convolutional layer is $M \times K_W \times K_H/S$: where

M = number of filters (equal to number of channels

K_H = kernel support height

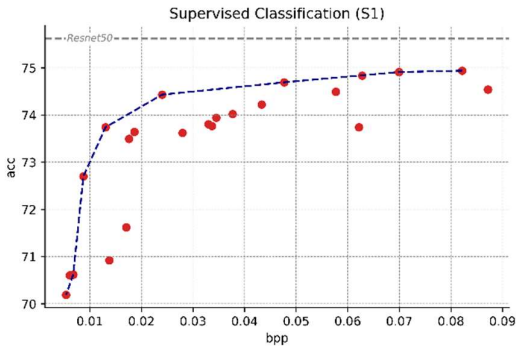
K_W = kernel support width and

S = down- or upsampling stride, where $\uparrow$ indicates upsampling and $\downarrow$ downsampling.

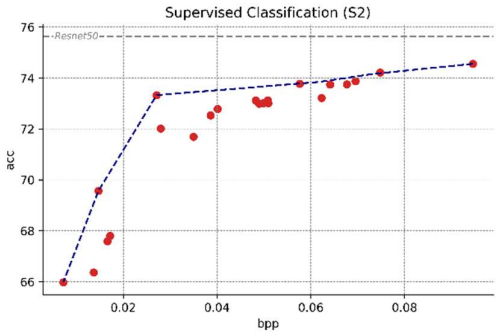
Complexity, Architecture and Task-performance of Various Bottleneck Topologies

Bottleneck Topology	Classification Split Point: layer4[2]				Segmentation Split Point: Resnet50.layer4[2]			
	Acc	BPP	# Param	Architecture detail of encoder	mIoU	BPP	# Param	Architecture detail of encoder
One conv layer	73.37	0.068	4.72M	(128X3X3/2↓)	0.666	0.093	4.46M	(64X3X3/4↓)
Two conv layers	72.79	0.069	5.58M	(2048X3X3/2↓) (128X1X1/1↓)	0.660	0.086	45.36M	(2048X3X3/4↓) (64X1X1/1↓)
One depthwise sep conv layer	74.91	0.069	0.57M	(128X3X3/2↓)	0.671	0.089	0.31M	(64X3X3/4↓)
Two depthwise sep conv layers	74.58	0.068	0.58M	(2048X3X3/2↓) (128X1X1/1↓)	0.638	0.102	0.32M	(2048X3X3/4↓) (64X1X1/1↓)

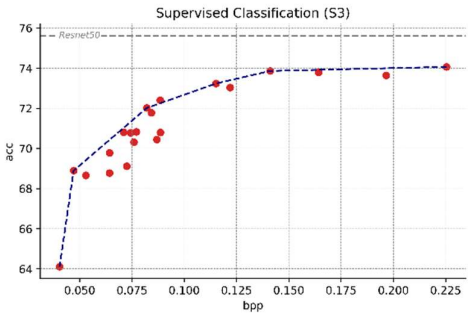
Supervised



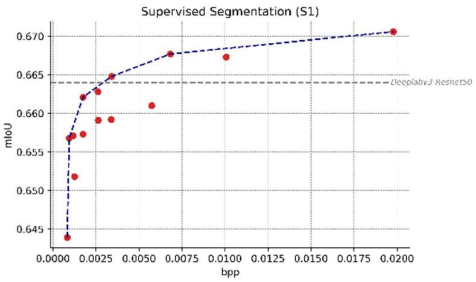
Lambda	Out channel	Quantization Factor	Stride	Accuracy	BPP
1.27E-08	96	0.88687	2	74.938	0.0821
1.21E-08	128	3.38315	2	74.91	0.0699
1E-09	96	8	2	74.692	0.0477
1.00E-08	96	1.5	2	74.43	0.024
3.71E-08	96	6.35293	4	73.74	0.0130
8.86E-08	64	7.42024	4	72.698	0.0087
7.97E-07	32	5.58587	4	70.618	0.0067
6.85E-08	32	7.01792	4	70.188	0.0054



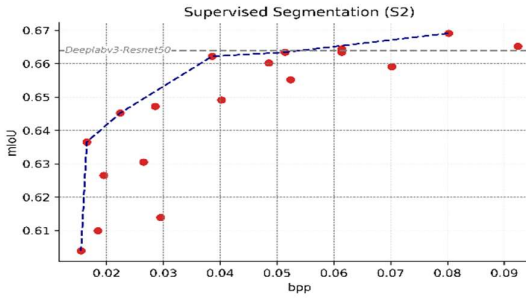
Lambda	Out channel	Quantization Factor	Stride	Accuracy	BPP
1.00E-09	128	1.345	2	74.55	0.0946
3.44E-08	128	3.87209	2	74.15	0.0732
8.46E-07	96	1.88755	2	73.839	0.0618
4.30E-06	64	3.20017	2	73.32	0.0271
3.98E-08	96	3.9559	4	69.568	0.0147
2.18E-07	128	3.85557	4	65.98	0.00719



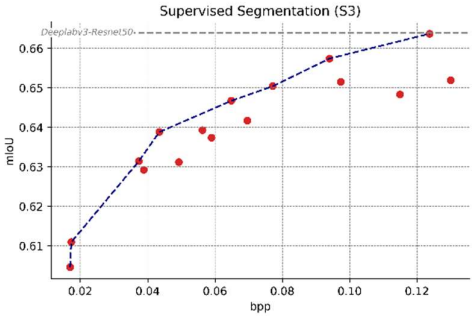
Lambda	Out channel	Quantization Factor	Stride	Accuracy	BPP
9.72E-07	64	0.51635	2	74.058	0.2254
1E-08	64	3.66894	2	73.854	0.1410
1.13E-07	64	7.54644	2	73.228	0.1151
8.25E-08	64	6.94574	2	72.00999	0.0821
5.61E-08	96	7.26885	4	68.892	0.0471
7.54E-08	32	0.50803	4	64.112	0.0403



Lambda	Out channel	Quantization Factor	Stride	mIoU	BPP
3.62E-03	64	3.245	4	0.6691	0.0802
8.76E-03	32	0.512	4	0.6634	0.0514
9.52E-04	32	1	4	0.6621	0.0385
6.79E-04	32	1.892	6	0.6452	0.0224
2.35E-03	32	4.782	4	0.6365	0.0165
1.67E-02	32	5.432	6	0.6039	0.0155

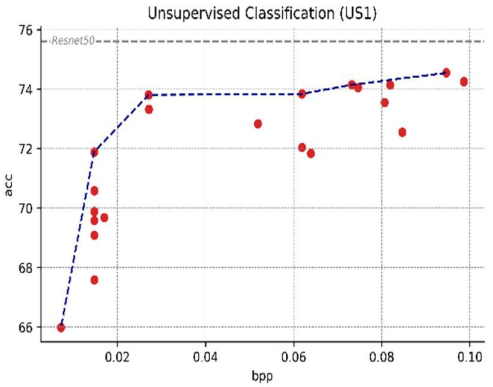


Lambda	Out channel	Quantization Factor	Stride	mIoU	BPP
0.0025	16	1.278	6	0.6706	0.01978
0.0006	16	4.23	6	0.6677	0.00682
0.082	8	0.625	6	0.6648	0.00341
0.0001	4	3.158	6	0.6621	0.00175
0.0004	2	2.14	6	0.6568	0.00094
0.076	2	0.834	6	0.6439	0.00082

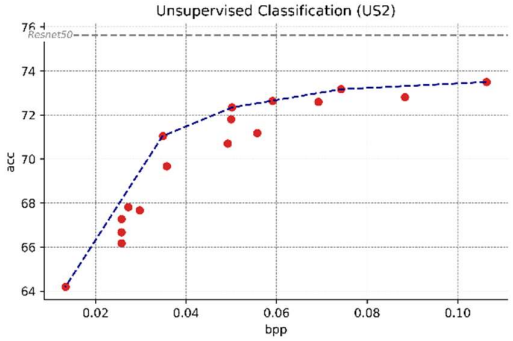


Lambda	Out channel	Quantization Factor	Stride	mIoU	BPP
0.00097	64	1.1749	2	0.6637	0.1237
0.0001	64	1.167	4	0.6574	0.094
0.00060	64	1.42407	4	0.6504	0.0771
0.00026	48	1.18992	4	0.6467	0.0648
0.0001	32	4	4	0.6388	0.0434
0.00076	32	2.88061	4	0.6314	0.0373
0.00783	32	0.5	6	0.6109	0.0173
0.00833	32	0.72141	6	0.6045	0.0169

Unsupervised

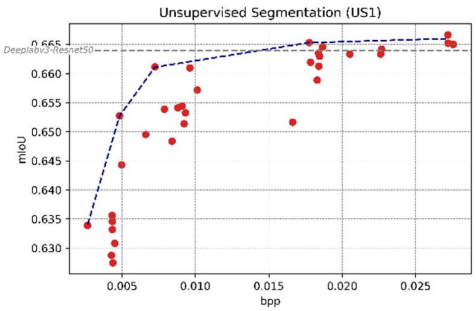


Lambda	Out channel	Quantization Factor	Stride	Accuracy	BPP
1.00E-07	128	0.5	2	74.55	0.0946
7.23E-10	128	1.5	2	74.15	0.0732
3.20E-08	96	3.062	2	73.839	0.0618
1.72E-07	128	1	4	73.81	0.0271
1.00E-07	64	2.78	4	71.89	0.0147
1.49E-06	64	3.3	4	65.98	0.0071

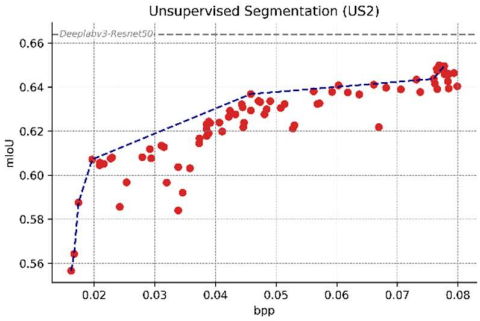


Lambda	Out channel	Quantization Factor	Stride	Accuracy	BPP
1.00E-08	128	1	2	73.498	0.1063
1.00E-07	128	2.623	2	73.17	0.0742
4.65E-08	96	4.37	2	72.34	0.0501
5.75E-09	64	5.33	2	71.04	0.0348
4.65E-10	96	2.73	4	64.18	0.0133

Lambda	Out channel	Quantization Factor	Stride	mIoU	BPP



1.00E-08	32	1	6	0.666	0.0271
1.00E-08	16	2.56	6	0.6653	0.0177
1.00E-06	12	3.127	6	0.6611	0.00724
1.00E-06	8	0.893	6	0.6527	0.00483
1.00E-06	4	1.39	6	0.6339	0.00264



Lambda	Out channel	Quantization Factor	Stride	mIoU	BPP
1.197E-12	48	1.904	4	0.649	0.07771
1.57E-13	48	1.995	4	0.643	0.0759
1.14E-12	48	1.904	4	0.636	0.0458
1E-16	32	3.48655	6	0.607	0.0196
0.000001	32	4.61145	6	0.587	0.0174
0.000001	32	5.27286	6	0.564	0.0166
1.54E-10	32	6	6	0.556	0.0162