On the iterative refinement of densely connected representation levels for semantic segmentation

Arantxa Casanova[†‡] Guillem Cucurull[†‡] Michal Drozdzal[†*] Adriana Romero[†*] Yoshua Bengio[†]

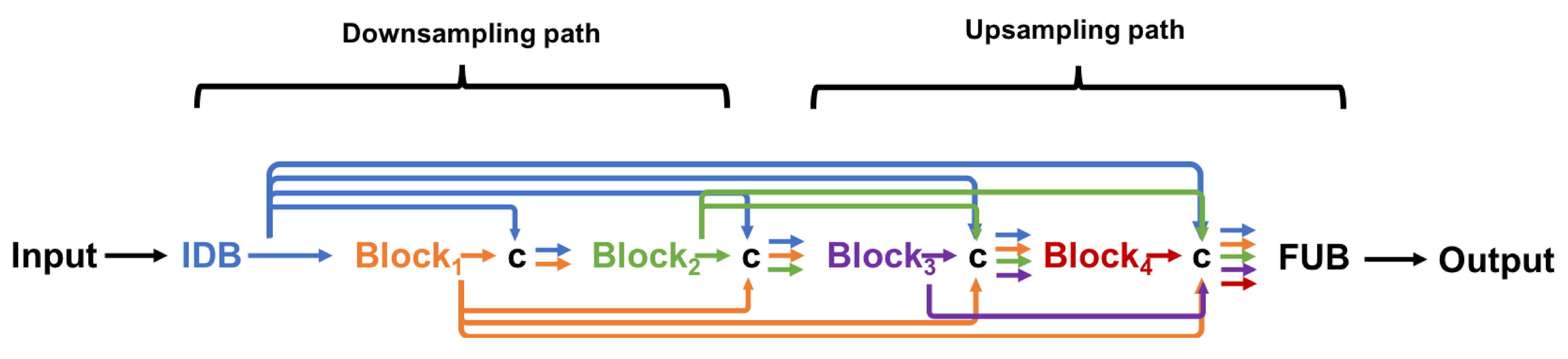
[†]*Montreal Institute for Learning Algorithms* [‡] *Computer Vision Center* [*] *Facebook AI Research*

github.com/ArantxaCasanova/fc-drn

Overview

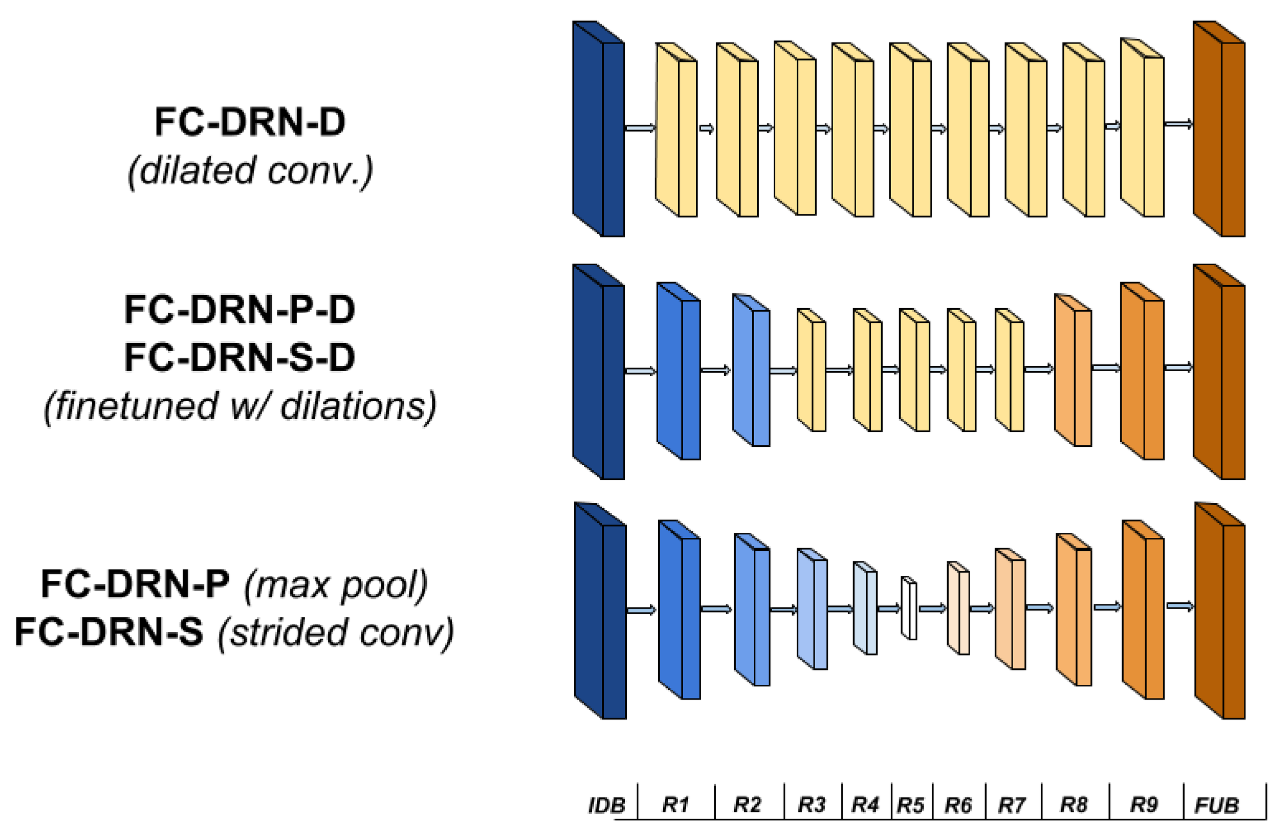
- Fully Convolucional Dense-ResNet (FC-DRN) is an architecture which exploits the benefits of both ResNets [1] & DenseNets [2], by densely connecting ResNet modules, which iteratively refine features at the same level of abstraction.
- We study the differences introduced by distinct receptive field enlargement methods and their impact on the performance of FC-DRN. We observe that:
  - Downsampling operations outperform dilations when trained from scratch.
  - Dilated convolutions are useful during the finetuning step of the model.
  - Coarser representations require less refinement steps.
  - ResNets are good regularizers: they reduce model capacity when needed.
- We report state-of-the-art results on Camvid [3] with at least 2x fewer parameters than existing methods.

Fully Convolutional DenseResNet (FC-DRN)



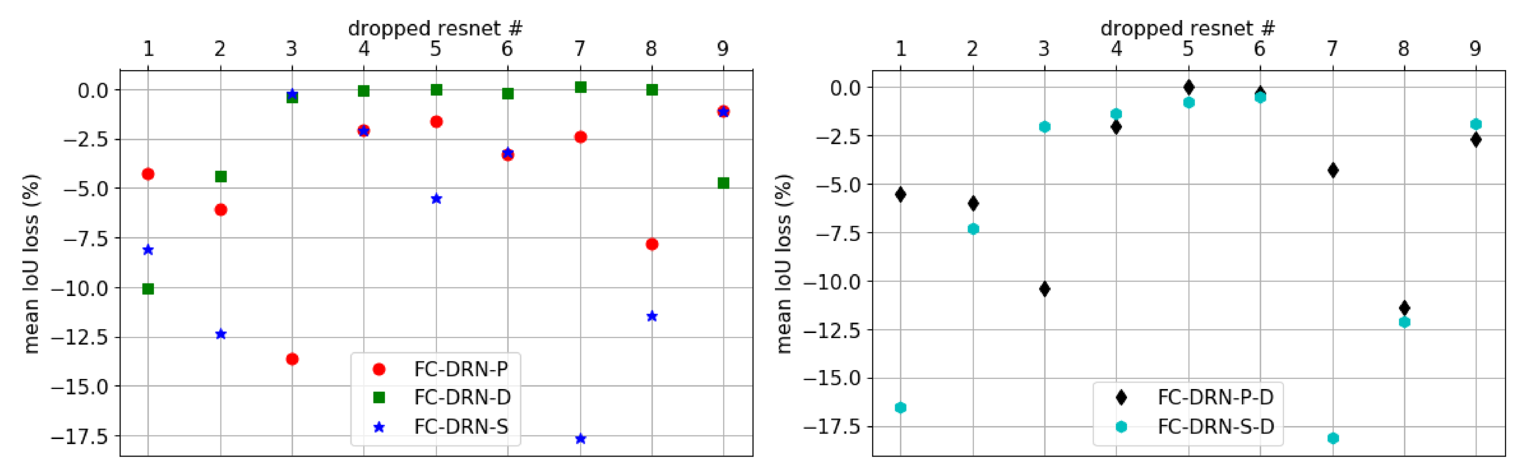
IDB: Initial Downsampling Block **FUB**: Final Upsampling Block **Block**: Convolution + ResNet

FC-DRN Transformation variants

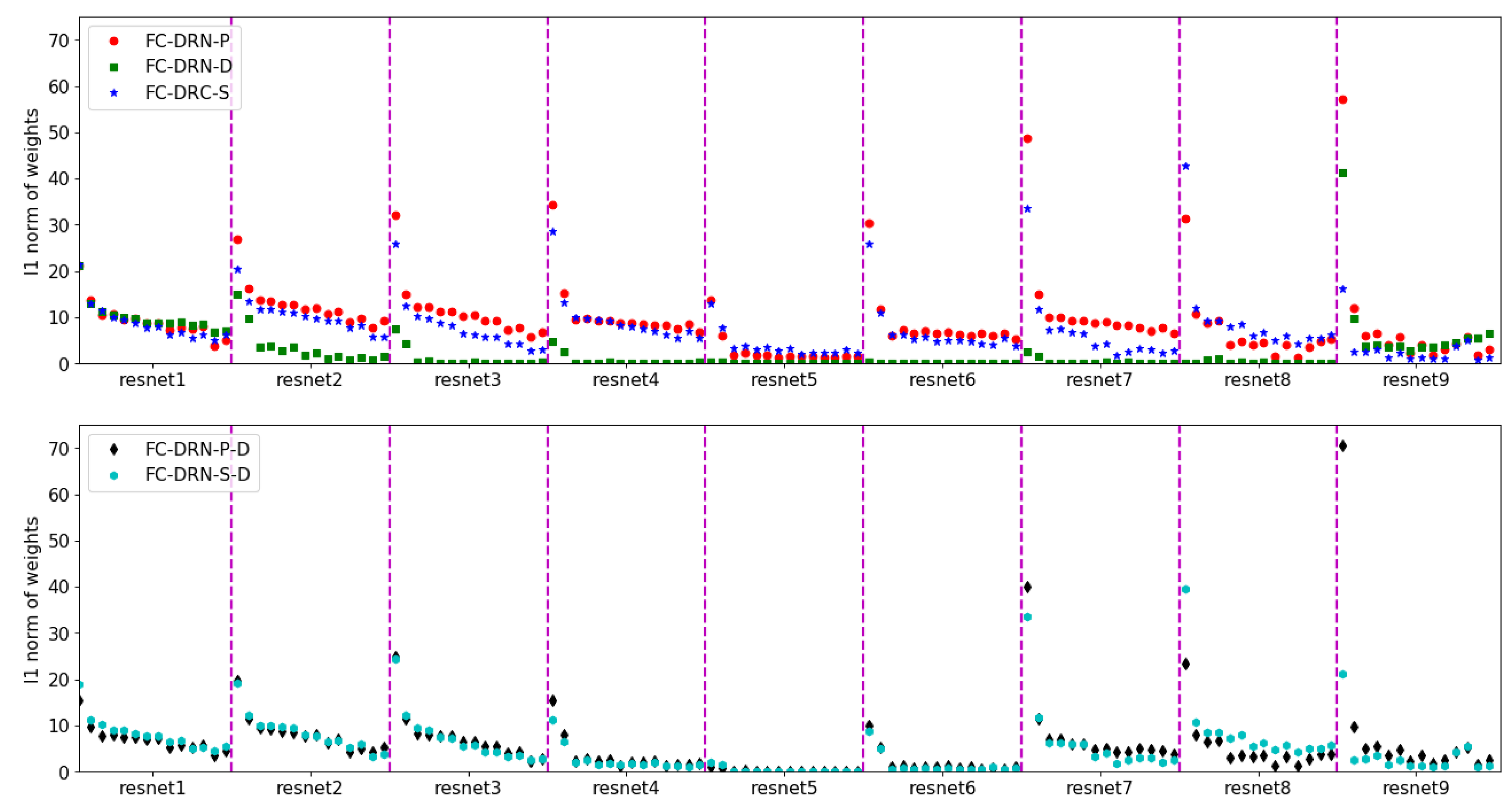


Analysis of trained models

Effect of dropping ResNets on performance



Visualizing the norm of the weights



Observations

- In FC-DRN-P/S, removing ResNets 4-6 mildly affects performance.
- In FC-DRN-D, the capacity of ResNets 4-8 is not used in a trained model, since the weight norms are very small.
- In general, removing the layers with small weight norm from a trained model only slightly affects the performance.
- Finetuning with dilations reduces the weight norms, especially in the layers close to network's bottleneck.

References

[1] K. He et al. Deep residual learning for image recognition. *CVPR*, 2016.

[2] G. Huang et al. Densely connected convolutional networks. *CoRR*, 2016.

[3] Gabriel J et al. Brostow. Segmentation and recognition using structure from motion point clouds. In *ECCV*. Springer, 2008.

[4] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *CVPR*, 2016.

[5] S. Jégou et al. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *CVVT, CVPRW*, 2017.

[6] Md Amirul et al. Islam. Gated feedback refinement network for dense image labeling. In *CVPR*, 2017.

Experiments

Comparison of different FC-DRN variants

Architecture	Val. IoU (%)	Val. acc (%)	mean IoU loss [%]	compression rate
FC-DRN-P	**81.1**	**96.1**	-1.6	1.08
FC-DRN-S	80.3	95.9	-5.4	1.08
FC-DRN-D	77.4	95.5	-0.8	1.38
FC-DRN-P-D	**81.7**	**96.0**	-1.0	1.15
FC-DRN-S-D	81.1	96.0	-1.7	1.15

Comparison to SotA on Camvid dataset

Model	Test IoU(%)	Test accuracy(%)	# params
Dilation8 + FSO [4]	66.1	88.3	130
FC-DenseNet67 [5]	65.8	90.8	3.5
FC-DenseNet103 [5]	66.9	91.5	9.4
G-FRNet [6]	68.0	90.8	30
FC-DRN-P-D	68.3	91.4	3.9
FC-DRN-P-D (+soft T.)	**69.4**	**91.6**	3.9

Qualitative results

Test image [3] Test ground truth [3] **FC-DRN-P-D**

Conclusions

- We highlighted the potential of FC-DRN achieving state-of-the-art on Camvid, with at least 2x fewer parameters.
- We analyzed different downsampling operations and carefully inspected each model, showing that:
  1. ResNets are good regularizers: they reduce model capacity when needed.
  2. Coarser representations require less refinement steps.
  3. Pooling generalizes better, while the benefits of dilations only apply when combined with pre-trained networks that contain downsampling operations.