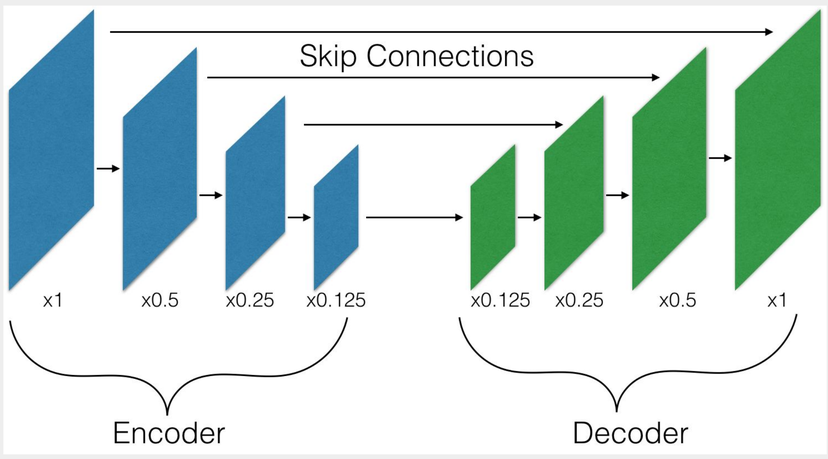
The Devil is in the Decoder

Zbigniew Wojna, Vittorio Ferrari, Sergio Guadarrama, Alireza Fathi,
Nathan Silberman, Liang Chieh Chen, Jasper Uijlings
zbigniewwojna@gmail.com, vittoferrari@google.com, sguada@google.com, nsilberman@google.com,
lcchen@google.com, alirezafathi@google.com, jrru@google.com

Problem statement:

- Encoder-decoder CNN is used in many vision problems that require per-pixel prediction.
- Different layers achieved superior performance on different problems [1,2,3].
- The choice of upsampling layer and skip connections were only studied in the context of a single problem.

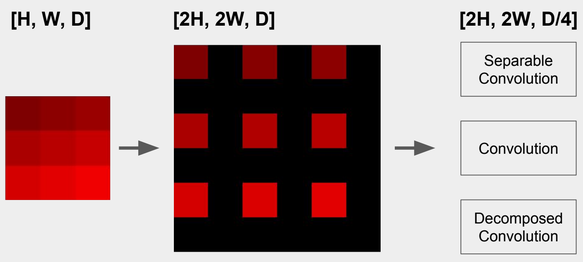


Upsampling method	# of parameters	# of operations	Comments
Transposed	$whIO$	$whWHIO$	
Dec. Transposed	$(w+h)IO$	$(w+h)WHIO$	Subset of Transposed
Sep. Transposed	$whI+IO$	$whWHI+WHIO$	Subset of Transposed
Conv and Depth To Space	$whI(4O)$	$whWHI(4O)$	
Bilinear with Conv	$whIO$	$wh(2W)(2H)IO$	
Bilinear additive with Conv	$whIO$	$wh(2W)(2H)(I/4)O$	

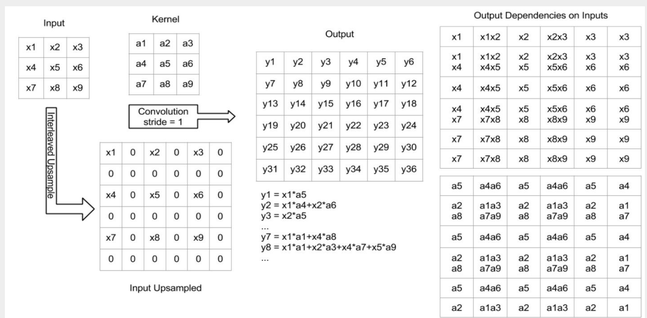
References

1) Laina, Iro, et al. “Deeper Depth Prediction with Fully Convolutional Residual Networks”
2) Shi, Wenzhe, et al. "Is the deconvolution layer the same as a convolutional layer?”
3) Odena, Augustus, Vincent Dumoulin, and Chris Olah. "Deconvolution and checkerboard artifacts."

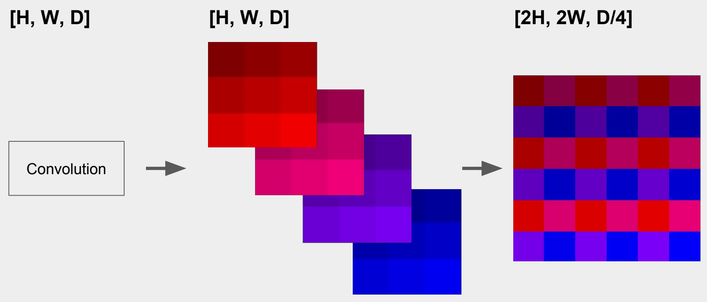
Transposed Convolution



Transposed Convolution



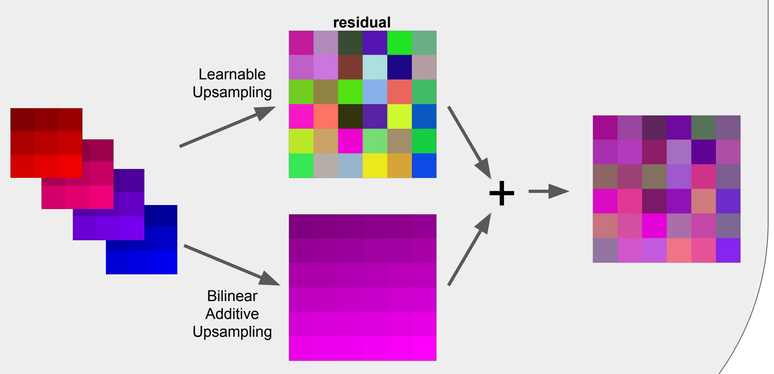
Conv + Depth To Space



Bilinear Upsampling + Conv

Contribution: Bilinear Additive Upsampling

Contribution: Residual-like connections



Conclusions:

1) Decoders matter, high variance of results.

2) Residual connection helps.

3) Additive linear upsampling gives the best overall results.

Problem	Residual	Skip	Semantic Segmentation	Depth Prediction	Colorization	Super-resolution	Instance Edge Detection
Dataset			VOC Pascal 2012	NYUv2 Depth	ImageNet	CelebA	VOC Pascal 2012
Metric			mIoU	MRE	AUC	SSIM	f-measure
Transposed	N	N	0.651	0.162	0.951	0.674	0.248
Transposed	N	Y	0.659	0.165	0.954		0.639
Decomposed Transposed	N	N	0.642	0.164	0.951	0.68	0.458
Decomposed Transposed	N	Y	0.652	0.166	0.953		0.522
Separable Transposed	N	N	0.659	0.163	0.952	0.676	0.57
Separable Transposed	N	Y	0.671	0.164	0.948		0.57
Upsample Bilinear + Conv	N	N	0.62	0.198	0.949	0.593	0.451
Upsample Bilinear + Conv	N	Y	0.656	0.174	0.949		0.565
Conv + Depth To Space	N	N	0.649	0.162	0.95	0.596	0.5
Conv + Depth To Space	N	Y	0.644	0.174	0.953		0.647
Bilinear Additive Upsampling + Conv	N	N	0.661	0.165	0.949	0.594	0.619
Bilinear Additive Upsampling + Conv	N	Y	0.669	0.169	0.952		0.653
Transposed	Y	N	0.655	0.164	0.951	0.686	0.622
Transposed	Y	Y	0.659	0.164	0.953		0.295
Decomposed Transposed	Y	N	0.645	0.171	0.954	0.682	0.243
Decomposed Transposed	Y	Y	0.637	0.162	0.951		0.531
Separable Transposed	Y	N	0.669	0.166	0.945	0.683	0.61
Separable Transposed	Y	Y	0.652	0.165	0.946		0.517
Upsample Bilinear + Conv	Y	N	0.653	0.171	0.949	0.684	0.53
Upsample Bilinear + Conv	Y	Y	0.651	0.175	0.954		0.537
Conv + Depth To Space	Y	N	0.652	0.174	0.953	0.686	0.624
Conv + Depth To Space	Y	Y	0.653	0.17	0.944		0.62
Bilinear Additive Upsampling + Conv	Y	N	0.658	0.165	0.951	0.683	0.625
Bilinear Additive Upsampling + Conv	Y	Y	0.654	0.167	0.952		0.643