Stacked Encoder-Decoder Networks for Human Semantic Segmentation

Owen J. Wang
Stanford University, Department of Computer Science

Abstract

Semantic segmentation is the problem of classifying regions of an image at [sub]pixel-level resolution to two or more predefined classes. In this project, we tackle the segmentation of human subjects in an image from their background by establishing a new CNN architecture, nicknamed **Diamondback**. Inspired by the work of Fu et. al. (2017) on stacked deconvolutional networks to approach semantic segmentation, we build off of their architecture to create a more lightweight model with about 25% the number of parameters and comparable performance.

Motivation

Improving semantic segmentation has numerous applications:
- Robotics: drone navigation, pedestrian detection
- Fashion / Clothing parsing
- Graphics: animation and CGI
- Medical image processing (3D segmentation of MRIs)

Balance speed and quality: lightweight model will help with time it takes to be practical with inference
- Real-time inference matters in autonomous vehicles
- AR fashion: projecting outfits, slow = bad customer experience

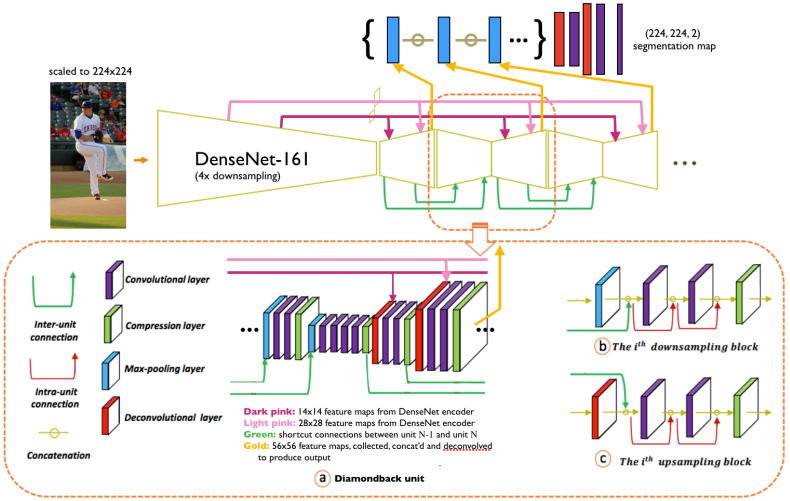
Related Work

- **DenseNet** (https://arxiv.org/abs/1608.06993)
  - Introduced concept of dense connections in CNNs, learning new features from concatenation of all previous features
- **FCN for Semantic Segmentation** (https://arxiv.org/pdf/1605.06211.pdf)
  - One of the first approaches to use CNNs for segmentation. Convolutions and max pooling to learn features, then upsampled 32x, 16x, and 8x resolutions to see which is the best.
- **The 100 Layers Tiramisu - FC DenseNets for Segmentation** (https://arxiv.org/pdf/1611.09326.pdf)
  - Encoder-decoder network with dense + shortcut connections, but only passed on newly learned features at each dense block for efficiency
- **Stacked Deconvolutional Networks for Segmentation** (https://arxiv.org/pdf/1708.04943.pdf)
  - Used dense connections, and a flexible number of encoder-decoder units with tons of shortcut connections to ease learning
- **Semantic Segmentation with Adversarial Networks** (https://arxiv.org/pdf/1611.08408.pdf)
  - Trained CNN for segmentation with an adversarial loss, reminiscent of GANs
- **DeepLab v3: Rethinking Atrous Convolutions** (https://arxiv.org/abs/1706.05587)
  - Used atrous convolutions of different scales and strides, to learn high- and low-level image context

Dataset

- MS COCO, 2017 split
- extracted only images and segmentation masks with people
- **64115** train, **2693** validation
- Did not have time to process data and evaluate model on test (21K samples)

Model Architecture & Details



Training
- Adam optimizer, initial learning rate 1e-4, decreased x4 at epoch 14.
- Dropout(0.2) on all convolution and deconvolution layers, 1e-4 L2 regularization.
- Data augmentation: horizontal flip, +/-20° rotation and up to 10% translation.
- ~6 hr 40 min per epoch.

Implementation Details
- Convolutional layer: composition of BatchNorm, 3x3 convolution, ReLU
  - Convolutions always learn 32 new filters
- Deconvolutional layer: composition of BatchNorm, 4x4 (stride 2) Transposed conv., ReLU
  - Deconvolutions produce half the number of filters as the input tensor has
- Number of compression layer filters in a unit (in order): 384, 512, 384, 256
- Number filters in final upsampling layers (before segm. map) is held constant at half total # concat'd filters

Results

Below are some predictions taken from the validation set. This subset of images was selected to demonstrate specific strengths and weaknesses of the model (see cont), but we sampled from a larger cross-section of the dev set to ensure they were representative. A deeper dive is on the right.

Fig 1. Sample dev images, predictions, and segmentation labels.

Results (cont)

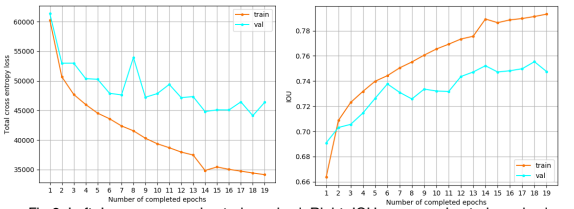


Fig 2. Left: Loss over epochs, train and val. Right: IOU over epochs, train and val. More training would likely have yielded further improvement.

Discussion on Predicted Samples

The sample predictions were produced by our Diamondback M2 model trained for 18 epochs, as this showed the highest val IOU and lowest val loss. We point out some worthy observations from DM2's predictions, and cite relevant samples in parentheses.

Smooth outlines of humans: Predictions have smoother, tighter-fit outlines of people than ground-truth labels even provide. (1,3)

Detects people at varying scales: Larger and smaller subjects in the image are both segmented successfully. (2,4,5)

Ignores image noise: Noise in the image such as watermarks are successfully filtered out. (2)

Mixed results with occlusions: Model learns to avoid irrelevant objects occluding people at times, but fails to do so at smaller scales / low-contrast regions. (1,4,5)

Fails to preserve masks' continuity in low-contrast regions: The model has a harder time generating a smooth, accurate segmentation when the image has low contrast in color/shade; see the old man's foot. (3)

Loses fine-grained detail at small scales: Fine-grained details like a hole formed by someone's arm-elbow or distinction between a faraway person's legs are lost. (1,2)

Quantitative Results

Paper	Evaluation dataset	mIOU
FCN (2015)	VOC2012	62.7
100 Tiramisu (2016)	CamVid, Gatech	66.9, 79.4
Adversarial Networks (2016)	VOC2012	72.0
SDN (2017)	COCO + VOC2012	79.2 **86.6** (COCO pretraining)
DeepLab v3 (2017)	COCO + VOC2012	79.7 85.7 (COCO pretraining)
Diamondback M2	COCO	75.5

Fig 3. Performance compared with other segmentation models. **Note that no comparison in this chart is perfect,** as other papers approach general segmentation tasks not solely focused on humans.

Future Directions

- Longer training time with more computational resources
- Swish activation for deeper networks
- include hierarchical supervision
- experiment with feature maps from atrous/dilated convolutions