JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

EN.601.482/682 Deep Learning

What does this network do?

Mathias Unberath, PhD
Assistant Professor
Dept of Computer Science
Johns Hopkins University

When I improve model accuracy by copying from Github an architecture I don't understand.

Fast R-CNN

- ROIs?
- Fast?
- Multi-task loss?
- Design choices?

Fast R-CNN

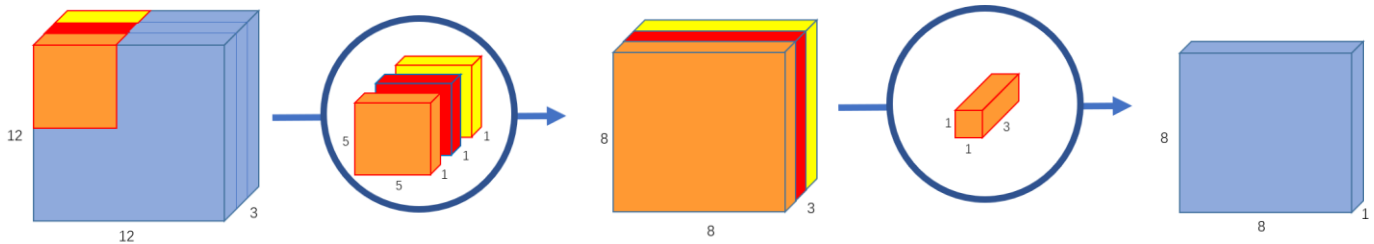
- FCN layers?
  - Constant size of region proposals → Bad proposals are classified as catch-all “background”
  - Remaining ROIs are pooled into constant size for compatibility with FCN
- Why is it fast?
  - Features are extracted only once for entire image, not for every ROI
- Multi-task loss?
  - Simultaneous learning of both classification and ROI BB refinement
  - Multiple losses: Multiple sources of information for SGD
- Design choices specific to object detection?
  - BB refinement is unique
  - Adaptive ROI alignment and pooling

DeepLabV3+

- Outperforming U-net?
- Atrous convolutions?
- Depth-wise separable convolutions?

DeepLabV3+

- Why could this outperform U-net?
  - Atrous convolutions to gather information on multiple scales during encoding
- Advantage of atrous convolutions?
  - Process information at multiple scales without additional parameters
- Depth-wise separable convolutions?
  - Filter kernel size k * k (3), input channels n_in (64), output channels n_out (128)
  - Regular convolution: (k * k * n_in) * n_out [3*3*64*128=73728]
  - Depthwise seperable: (k * k * n_in) + n_in * n_out [3*3*64 + 64*128=8768]



WaveNet

- Why atrous convolutions?
- Conditional WaveNet?

WaveNet

- Why are atrous convolutions helpful for this architecture?
  - Raw audio is very high frequency, at least 16,000 samples per second
  - Network requires very short inference times, but sufficiently long temporal context
  - Atrous convolutions allow for drastic receptive field increase without additional parameters

- Difference between WaveNet and conditional WaveNet?
  - WaveNet simply generates audio, no control over what is generated
  - Conditional WaveNets introduce additional latent variable (either constant for speaker ID or lower frequency for words, e.g. via embedding) → This allows for control over generation