JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Residual Stacked Hourglass Networks for Human Pose Estimation

Sun Jay Yoo, Kevin Peng, Manan Aggarwal

(601.482) Machine Learning: Deep Learning- Spring 2019

Instructors: Mathias Unberath, Jie Ying Wu, Cong Gao

Purpose

- Human pose estimation: the detection of human figures and estimating their key body joints in images and video
- Study, evaluate, and compare three existing architectures for 2D single human pose estimation and propose our own

Dataset

Microsoft COCO 2017 for Keypoint Detection

- Images taken from everyday scenes
- 91 object categories (80 available with annotations)
- 1.5M labeled instances and 328K images

Existing Architectures

DeepPose*

- CNN-like
  - 7-layer AlexNet backend with an extra final layer
  - Frontend trained using L2 loss instead of classification loss
- Uses cascaded regression to refine predictions from the previous stage
  - Subsequent regressors see higher resolution inputs and learn exact features

*Toshev and Szegedy (arXiv:1312.4659)

Chained Prediction*

- RNN-like
  - Each spatial prediction is used in linear combination with previous hidden state and output as inputs
  - Parameters not tied across timesteps
- CNNx: feature extraction
- CNNy: deconvolution/inception ("deception") to make spatial predictions

*Gkioxari et al. (arXiv:1605.02346)

Stacked Hourglass*

- UNet-like
  - Repeated phases of pooling layers followed by upsampling (hourglass)
  - Skip connections within each hourglass preserves spatial awareness
- Spatial information rerouted within hourglasses to reassess spatial information in a global context

*Newell et al. (arXiv:1603.06937)

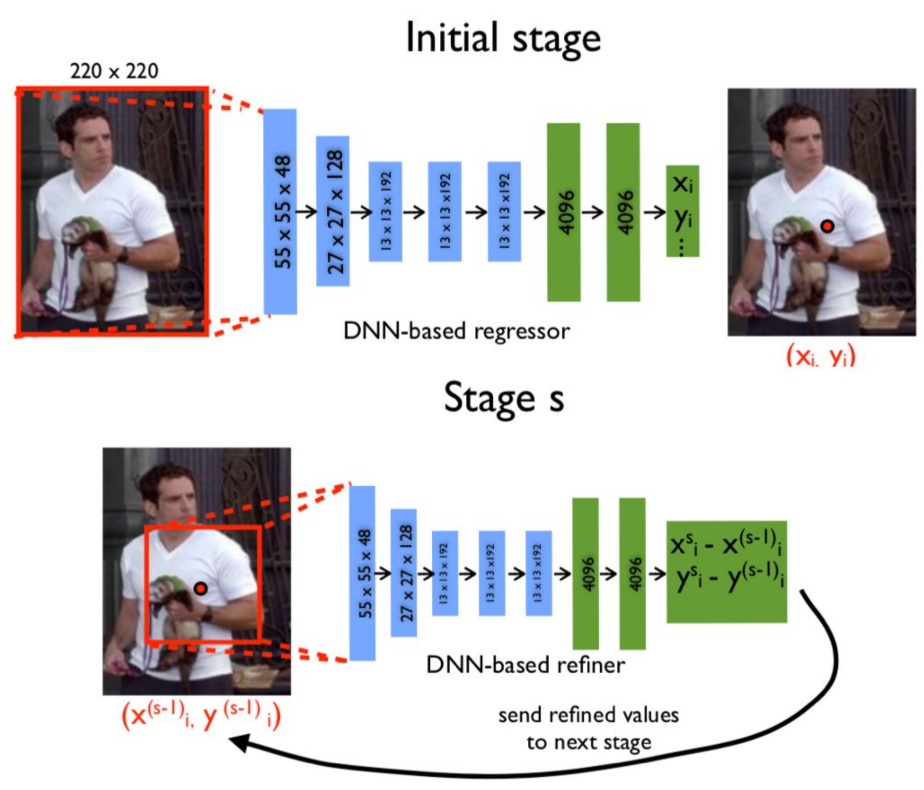

Figure 1: Overview of DeepPose architecture involving two stages. Top: schematic view of the DNN-based pose regression. Bottom: a cascaded regression applied on a sub-image to refine a prediction from the previous stage.

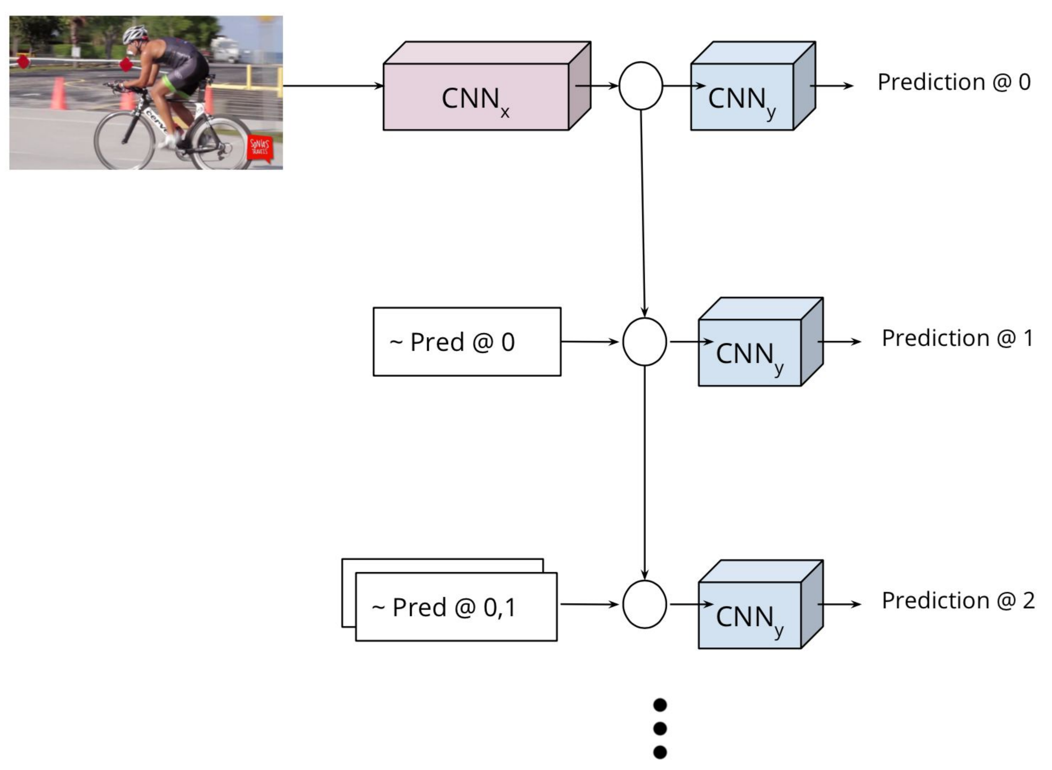

Figure 2: Visualization of the Chained Prediction model with the case of single images. Images are encoded with CNNx and decoder CNNy makes predictions using previous outputs and hidden states with a sequential model.

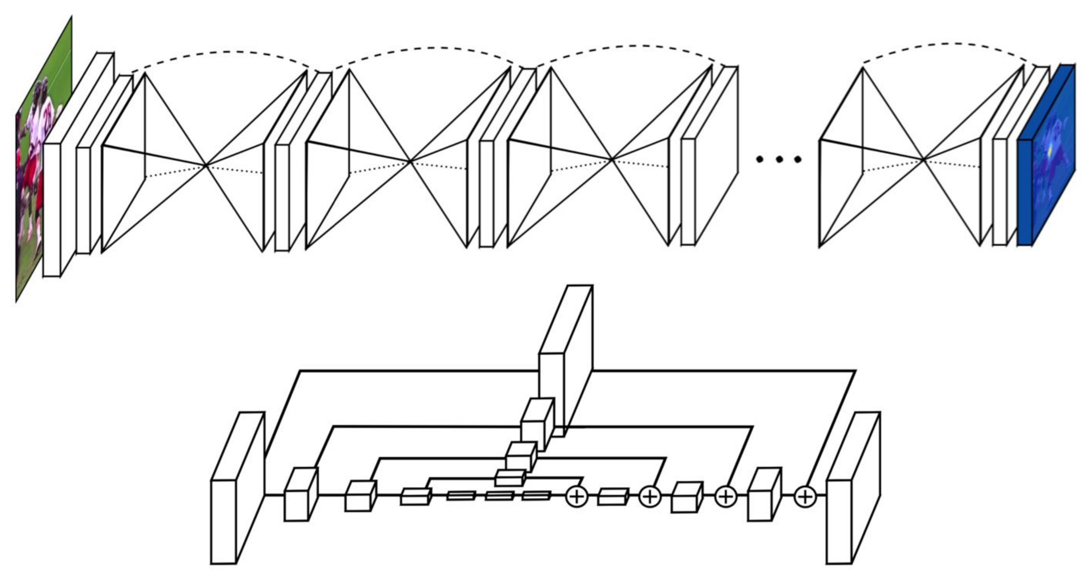

Figure 3: Top: overview of the Stacked Hourglass model with multiple hourglass modules. Bottom: single hourglass module with topological symmetry between routed highways.

Proposed Architecture

Residual Stacked Hourglass

- Combination of residual blocks from ResNet DeepPose with hourglass modules from Stacked Hourglass
  - Deeper network from three to five stacks
  - "Gradient highways" between symmetric hourglass modules
- Multi-scale information captured within each stack, multi-stage information captured between each stack

Methods

- Optimization via L2 loss and RMSProp with learning rate 2.5e-4
- Data augmentation
  - Translation: 2% of image width
  - Scaling: ±30% of image size
  - Rotations: ±40°
- Percentage Correct Keypoints (PCK) metric to evaluate precision
  - Percentage within normalized Gaussian of ground truth
- Trained on single NVIDIA Tesla P100

Results

Model	Parameters	PCK
DeepPose	40 M	70.4
Chained Prediction	26.5 M	82.0
Stacked Hourglass	12.6 M	84.7
Residual Stacked Hourglass	***31.1 M***	***81.1***

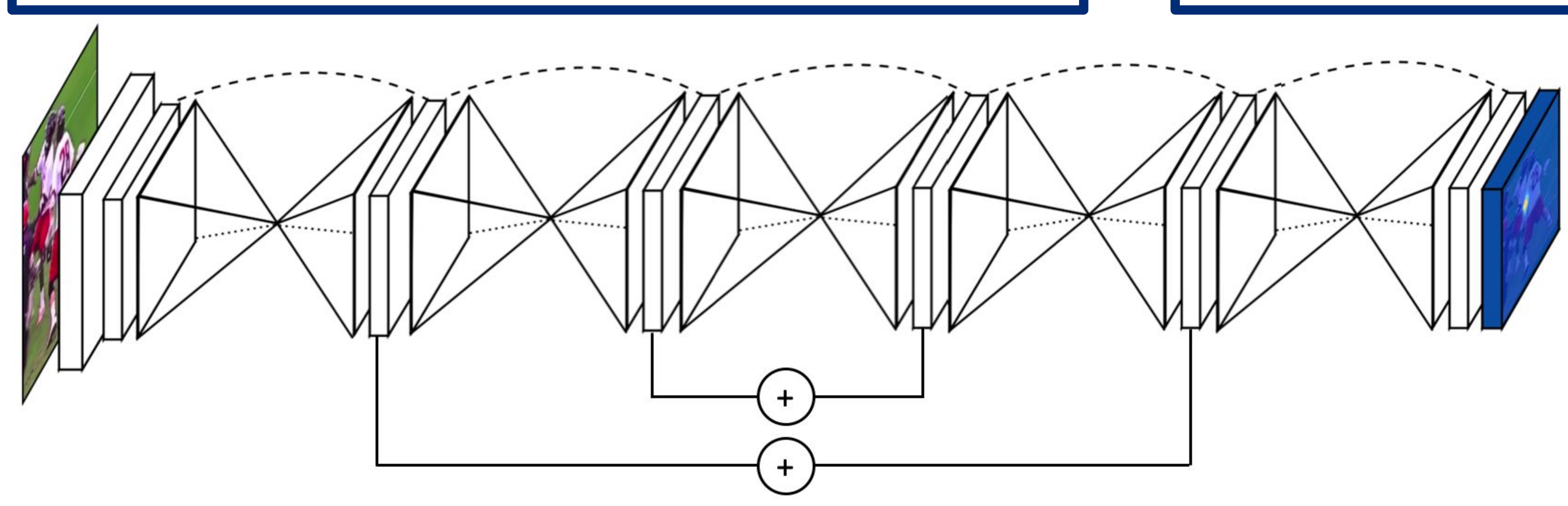

Figure 4: Schematic (adapted from the original Stack Hourglass paper) of the deep five-stack Residual Stacked Hourglass architecture with symmetric "gradient highways". Each hourglass remains the same as presented in Figure 3.

Figure 5: Example heatmap outputs produced by the network. The left image is the final pose estimate using the maximum activations across the sample heatmaps for each joint.

INTUITIVE SURGICAL®