Stacked Hourgloass
(2017)

ALejandro Newall et al.
Resume

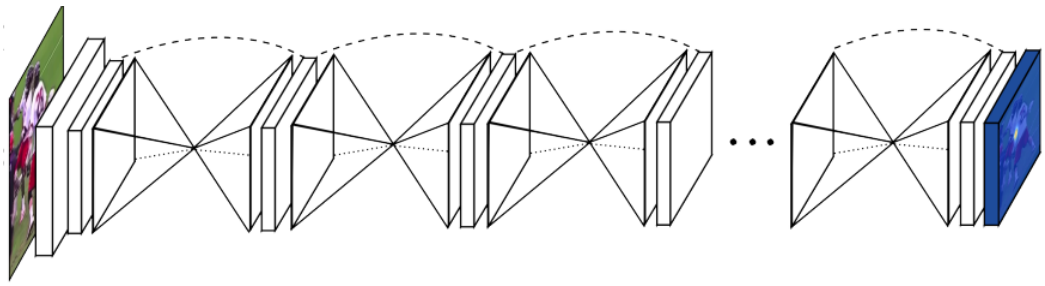
July 14, 2019

---

1 Introduction

In this work the authors propose a new architecture for human pose estimation, features are processed across all scales and consolidated to best capture the various spatial relationships associated with the body, the name hourglass comes from the subsequent pooling and upsamling, like many convolutionnal approaches that produces pixel wise outputs , the hourglass network pools down to a very low resolution and then upsamples and combines featurues across multiple resolutions, with a symetric topology.

2 The model



The hourglass module is caraterized by its symmetric nature and its design is motivated by the need to capture information at every scale, while local evidence is essential for identifying features likes faces and hands, a final pose estimate requires a coherent understanding of the full boady, the person's orientation, the relationships of adjacent joints, ect. and the hourglass network is capable of capturing all of these features and bring them together to output pixel wise prediction with a simple and minimal design.

The network contains skip layers to perserve the spatial information at each resolution, the network reaches the lowest resolution at 4x4 pixels (with an input of 256 and the first two downsamplings: conv with stride two and a max pool, and four max pools in the hourglass) allowing smaller spatial filers to be applied that compare feature across the whole image.

2.1 Architecture

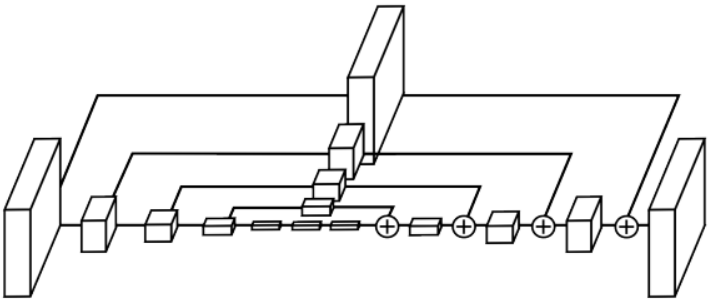
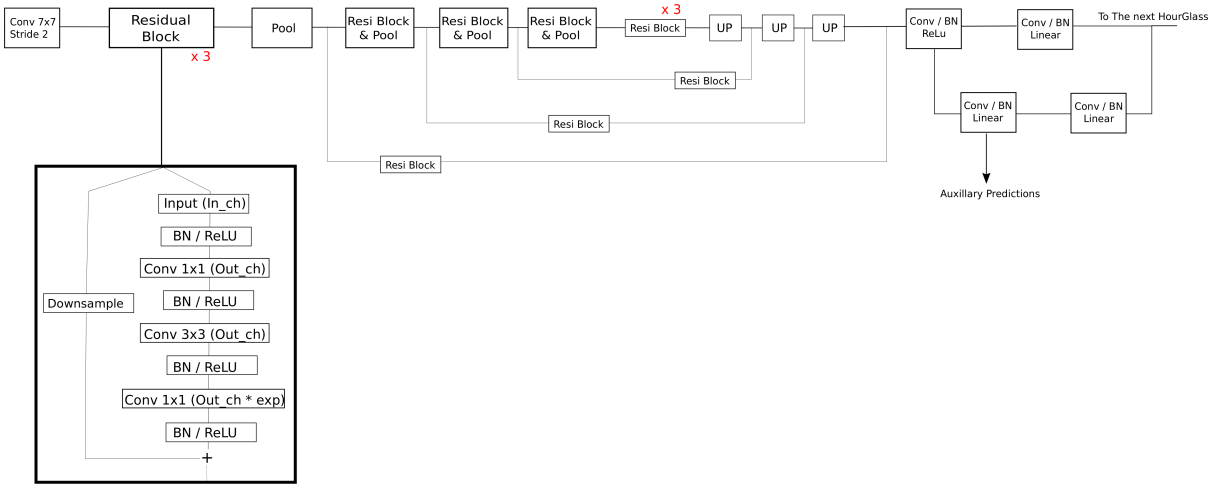


Fig. 3. An illustration of a single "hourglass" module. Each box in the figure corresponds to a residual module as seen in Figure 4. The number of features is consistent across the whole hourglass.

The hourglass is set up follows, convolutionnal and max pooling layers are used to process features down to a very low resolution, at each max pooling step, the network branches off and applied more convolutions (in the form of residual blocks) and applied a set of convolutions to the prepooled activations, and then the networl begins the top down sequence of upsampling and combination of features across scales unitil we reach the size of the input (in this case the size of the output is 64 x 64, so in_size / 4), after reaching the ourtput resolution of the network, two consecutive 1x1 convolutions are applied to produce the final network predictions, the output of the net is a set of heatmaps where for a given heatmap the network predicts the probability of a joint's presence at each and every pixel.

Here is a single hourglass with network:



In the paper, they stack the hourgloass parts (two), and use an intermediate supervision, the intermediate predictions are also added back to the main path as we see in the figure above.

3 Experimetns

The authors explore two main design choices in this work: the effect of stacking hourglass modules together, and the impact of intermediate supervision.

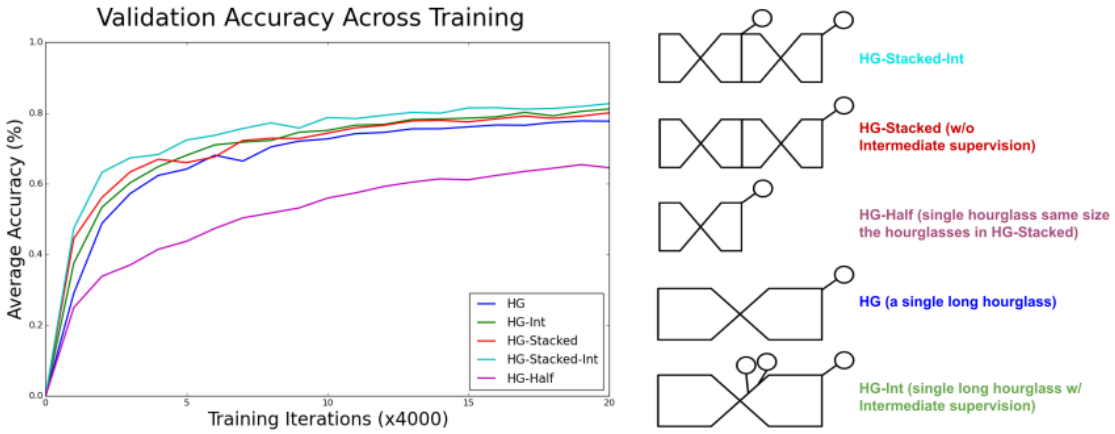


Fig. 8. Comparison of validation accuracy as training progresses. The accuracy is averaged across the wrists, elbows, knees, and ankles. The different network designs are illustrated on the right, the circle is used to indicate where a loss is applied

For the ablation study,They try to investegate the usage of a number of stacked hourglass networks, and how the refinnement process gives better resutls, using different number of hourglass network, they try to maintain the similar number of paramters of a faire comparaison:

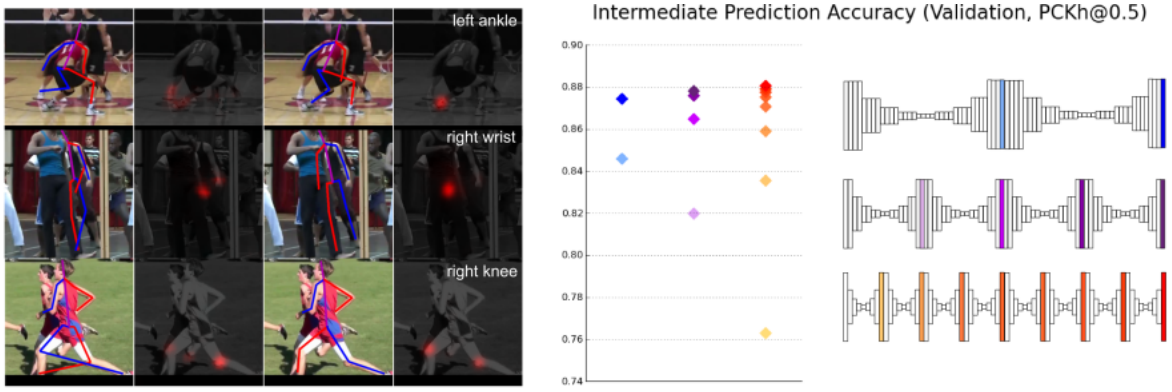


Fig. 9. Left: Example validation images illustrating the change in predictions from an intermediate stage (second hourglass) (left) to final predictions (eighth hourglass) (right). **Right:** Validation accuracy at intermediate stages of the network compared across different stacking arrangements.