Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians

Workshop on Object Understanding for Interaction 2015
Submission # ***

Convolutional neural nets (CNNs [13]) have demonstrated remarkable performance in recent history for visual tasks [12, 18, 20]. Such approaches compute hierarchical representations in a bottom-up, feedforward fashion. As biological evidence suggests [22], feedforward processing works effectively for vision at a glance tasks. However, vision with scrutiny tasks appear to require top-down feedback processing [8, 10], which is missing in the "uni-directional" CNNs. The main contribution of this work is to explore "bi-directional" architectures that are capable of feedback reasoning.

Feedback reasoning has played a central role in many classic computer vision models, such as hierarchical probabilistic models [9, 14, 24] and part-based models [3]. Interestingly, part-based model’s feed-forward inference can be written as a CNN [4], however the proposed mapping does not hold for feedback inference.

To endow CNNs with feedback inference, we treat neural units as non-negative latent variables in a quadratic energy function. When probabilistically normalized, our quadratic energy function corresponds to a Rectified Gaussian (RG) distribution, for which inference can be cast as a quadratic energy function. Inference on RGs can be unrolled into recurrent nets with rectified activations. Such architectures produce better features for “vision-with-scrutiny” tasks [7] (such as keypoint prediction) because lower-layers receive top-down feedback from above. Leg keypoints are much better localized with top-down knowledge (that may capture global constraints such as kinematic consistency).

Table 1: PCKh-0.5 on MPII-Val for QP1 and QP2 on a smaller network

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Body</td>
<td>59.7</td>
<td>59.6</td>
<td>50.7</td>
<td>61.4</td>
<td>58.7</td>
<td>60.9</td>
</tr>
<tr>
<td>Full Body</td>
<td>59.8</td>
<td>62.3</td>
<td>61.0</td>
<td>63.1</td>
<td>61.2</td>
<td>62.6</td>
</tr>
</tbody>
</table>

Figure 2: Illustration of unrolling two sequences of layer-wise coordinate updates into a recurrent net with skip connections.

Layerwise updates on Rectified Gaussian models

Feedforward neural nets

<table>
<thead>
<tr>
<th>Layerwise + Top-down</th>
<th>CNN</th>
<th>Recurrent-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>$h_2$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>$w_1$</td>
<td>$w_2$</td>
<td>$w_1$</td>
</tr>
<tr>
<td>$h_1$</td>
<td>$h_2$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>$w_1$</td>
<td>$w_2$</td>
<td>$w_1$</td>
</tr>
</tbody>
</table>

To demonstrate the benefits of integrating top-down feedback, we experimented with one-pass and two-pass RG variants of VGG-16 [18], which we refer to as QP1 and QP2. The architecture of unrolled QP2 is present in Fig. 3. We performed experiments on four challenging benchmark datasets of human faces and bodies, which are AFLW [11], COFW [2], Pascal Person [6], and MPII Human Pose [1].

On AFLW, we compared to ourselves for exploring best practices to build multi-scale predictors for facial keypoint localization. On COFW, our QP1 performs near the state-of-the-art, while QP2 significantly improves in accuracy of visible landmark localization and occlusion prediction. On Pascal Person, we show QP1 outperforms previous state-of-the-art by a large margin, while QP2 further improves accuracy by 2% without increasing model complexity. On MPII Human Pose, our QP2 model outperforms all prior work on localization accuracy over full-body keypoints. We present qualitative results in Fig. 4 and quantitative results in Table 1. As a side note, even visibility prediction is not in the standard evaluation protocol, we found QP2 outperforms QP1 on visibility prediction on both MPII Human Pose and Pascal Person dataset.

Given the consistent improvement of QP2 over QP1, we further explored QP2’s performance as a function of K. Due to memory limit, we trained a shallower network on MPII. As shown in Table 2, we concluded that: (1) all models with additional passes outperform the baseline QP1; (2) additional passes generally helps, but performance maxes out at QP4. A two-pass model (QP2) is surprisingly effective at capturing top-down info, while being fast and easy to train.


Figure 3: We show the architecture of QP₂ implemented in our experiments. QP₁ corresponds to the first half of unrolled QP₂, which essentially resembles the state-of-the-art VGG-16 CNN [18]. Note that QP₁ and QP₂ share the same number of parameters but differ in the number of layer-wise updates. Purple layers denote multi-scale predictors that predict keypoint heatmaps given activations from multiple layers. Multi-scale filters are efficiently implemented with coarse-to-fine upsampling [15], highlighted by the purple dotted rectangle. Dotted layers are layers having no effects on predictions in QP₂, hence not implemented to reduce memory.

Figure 4: Keypoint localization results of QP₂ on the MPII Human Pose testset. Our models are able to localize keypoints even under significant occlusions.


