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

Peiyun Hu[1], Deva Ramanan[2]
[1]UC Irvine [2]Carnegie Mellon University

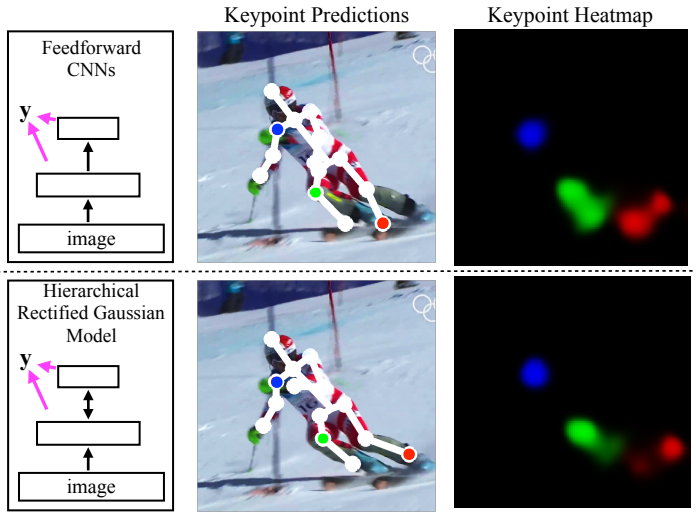

Figure 1: On the **top**, we show a state-of-the-art multi-scale feedforward net, trained for keypoint heatmap prediction, where the blue keypoint (the right shoulder) is visualized in the blue plane of the RGB heatmap. The ankle keypoint (red) is confused between left and right legs, and the knee (green) is poorly localized along the leg. We believe this confusion arises from bottom-up computations of neural activations in a feedforward network. On the **bottom**, we introduce hierarchical Rectified Gaussian (RG) models that incorporate top-down feedback by treating neural units as latent variables in 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).

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 program (QP) [19]. The QP's coordinate descent optimization steps, as we demonstrated in the paper, can be "unrolled" into a recurrent neural net with rectified linear units. An illustration of unrolling two sequences of coordinate updates is visualized in Fig. 2. This observation allows us to discriminatively-tune RGs with neural network toolboxes: *we tune Gaussian parameters such that, when latent variables are inferred from an image, the variables act as good features for discriminative tasks*.

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 QP_1 and QP_2. The architecture of unrolled QP_2 is present in

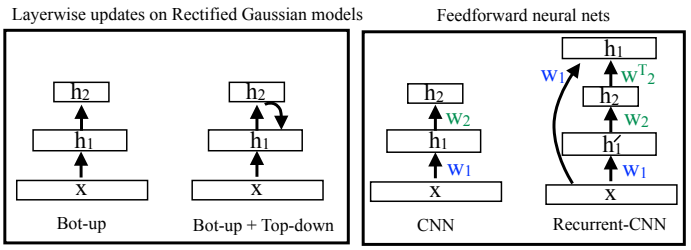

Figure 2: Illustration of unrolling two sequences of layer-wise coordinate updates into a recurrent net with skip connections.

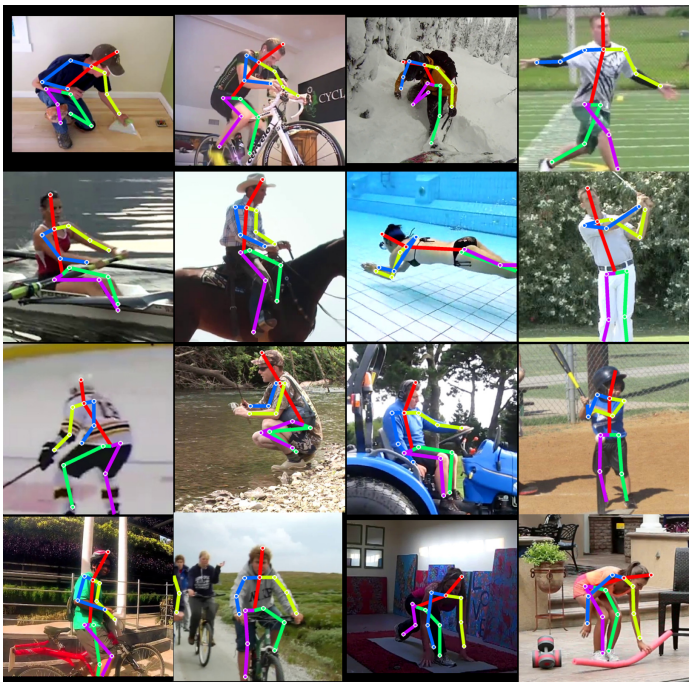

Figure 4: Keypoint localization results of QP_2 on the MPII Human Pose testset. Our models are able to localize keypoints even under significant occlusions.

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 QP_1 performs near the state-of-the-art, while QP_2 significantly improves in accuracy of visible landmark localization and occlusion prediction. On Pascal Person, we show QP_1 outperform previous state-of-the-art by a large margin, while QP_2 further improves accuracy by 2% without increasing model complexity. On MPII Human Pose, our QP_2 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 QP_2 outperforms QP_1 on visibility prediction on both MPII Human Pose and Pascal Person dataset.

Given the consistent improvement of QP_2 over QP_1, we further explored QP_k'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 QP_1; (2) additional passes generally helps, but performance maxes out at QP_4. A two-pass model (QP_2) is surprisingly effective at capturing top-down info, while

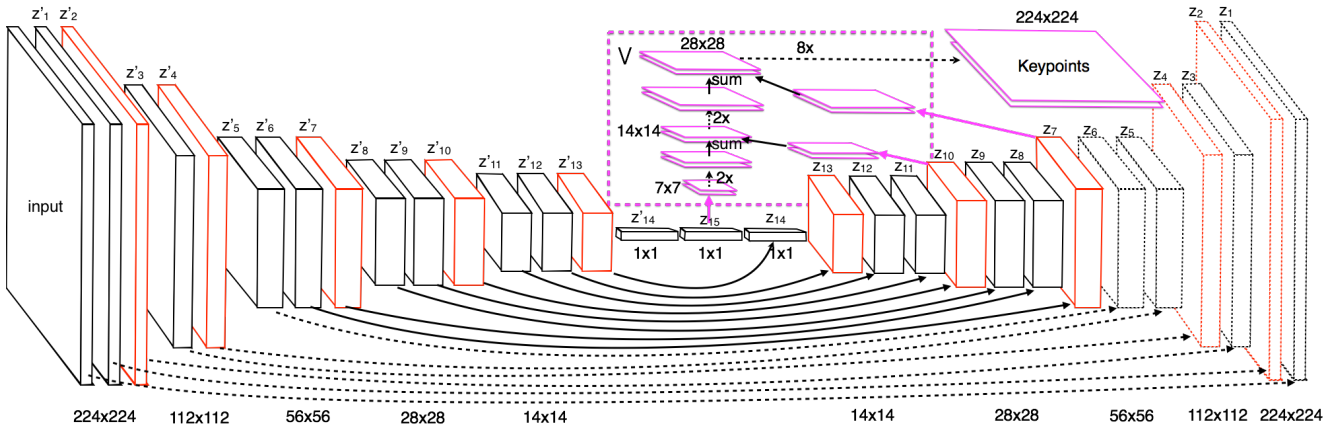

Figure 3: We show the architecture of QP_2 implemented in our experiments. QP_1 corresponds to the first half of unrolled QP_2, which essentially resembles the state-of-the-art VGG-16 CNN [18]. Note that QP_1 and QP_2 share the same number of parameters but differ in the number of layer-wise updates. 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 QP_2, hence not implemented to reduce memory.

	Head	Shou	Elb	Wri	Hip	Kne	Ank	Upp	Full
GM [5]	-	36.3	26.1	15.3	-	-	-	25.9	-
ST [17]	-	38.0	26.3	19.3	-	-	-	27.9	-
YR [23]	73.2	56.2	41.3	32.1	36.2	33.2	34.5	43.2	44.5
PS [16]	74.2	49.0	40.8	34.1	36.5	34.4	35.1	41.3	44.0
TB [21]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	**84.5**	82.0
QP_1	94.3	90.4	81.6	75.2	80.1	73.0	68.3	82.4	81.1
QP_2	95.0	91.6	83.0	76.6	81.9	74.5	69.5	83.8	**82.4**

Table 1: PCKh-0.5 on MPII-test using the recommended benchmark protocol[1].

K	1	2	3	4	5	6
Upper Body	57.8	59.6	58.7	**61.4**	58.7	60.9
Full Body	59.8	62.3	61.0	**63.1**	61.2	62.6

Table 2: PCKh-0.5 on MPII-Val for QP_k on a smaller network

being fast and easy to train.

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1513–1520. IEEE, 2013.

[3] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[4] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *CVPR*. IEEE, 2015.

[5] Georgia Gkioxari, Pablo Arbeláez, Lubomir Bourdev, and Jagannath Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR*, 2013.

[6] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[7] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002.

[8] Minami Ito and Charles D Gilbert. Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, 22 (3):593–604, 1999.

[9] Ya Jin and Stuart Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, 2006.

[10] Stephen M Kosslyn, William L Thompson, Irene J Kim, and Nathaniel M Alpert. Topographical representations of mental images in primary visual cortex. *Nature*, 378(6556):496–498, 1995.

[11] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, pages 2144–2151. IEEE, 2011.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*. IEEE, 2015.

[16] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *CVPR*, pages 588–595, 2013.

[17] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[19] Nicholas D Socci, Daniel D Lee, and H Sebastian Seung. The rectified gaussian distribution. *NIPS*, pages 350–356, 1998.

[20] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*. IEEE, June 2015.

[21] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christopher Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015.

[22] Rufin VanRullen and Simon J Thorpe. Is it a bird? is it a plane? ultra-rapid visual categorisation of natural and artifactual objects. *Perception-London*, 30(6):655–668, 2001.

[23] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE, 2011.

[24] Long Leo Zhu, Yuanhao Chen, and Alan Yuille. Recursive compositional models for vision: Description and review of recent work. *Journal of Mathematical Imaging and Vision*, 41(1-2):122–146, 2011.