Summaries for "Deeply-Supervised Nets"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author introduces support vector classifiers to each hidden layer and study the transparency of the intermediate layers to the overall classification, discriminativeness and robustness of learned features. They demonstrate fast convergence rate and more robustness of the proposed method. Experiments on four standard benchmark datasets verify the claims.

Main idea

To address the feature learning problem in deep learning architecture, the author adds SVM to each hidden layer and enforces direct and early supervision for both the hidden layers and the output layer. The assumption that local strong convexity of the optimization function is proposed by the writer. By forcing the whole convolution neural network to learn discriminate feature maps, the feature maps will become more discriminative. So the author states that this method will much more rapidly approach the region of good features than would be the case if only gradual backpropagation is used. Another point is that gradients vanishing or exploding may be avoided in this way.

Based on traditional CNNs, the author adds SVM to each layer then the overall objective function will be as follows:

$$\|\mathbf{w}^{(out)}\|^2 + \mathcal{L}(\mathsf{W}, \mathbf{w}^{(out)}) + \sum_{m=1}^{M-1} \alpha_m [\|\mathbf{w}^{(m)}\|^2 + \ell(\mathsf{W}, \mathbf{w}^{(m)}) - \gamma]_+,$$

where

$$\mathcal{L}(\mathsf{W}, \mathbf{w}^{(out)}) = \sum_{y_k \neq y} [1 - < \mathbf{w}^{(out)}, \phi(\mathbf{Z}^{(M)}, y) - \phi(\mathbf{Z}^{(M)}, y_k) >]_+^2$$

and

$$\ell(\mathsf{W}, \mathbf{w}^{(m)}) = \sum_{y_k \neq y} [1 - < \mathbf{w}^{(m)}, \phi(\mathbf{Z}^{(m)}, y) - \phi(\mathbf{Z}^{(m)}, y_k) >]_+^2$$

It can be written as F(W)=P(W)+Q(W) where P(W) means the first two terms and Q(W) means the last term.

Furthermore, the author demonstrates such three lemmas:

Lemma 1 $\forall m, m' = 1..M-1, and\ m' > m$ *if* $\|\mathbf{w}^{(m)}\|^2 + \ell((\hat{\mathsf{W}}^{(1)}, .., \hat{\mathsf{W}}^{(m)}), \mathbf{w}^{(m)}) \leq \gamma$ *then there exists* $(\hat{\mathsf{W}}^{(1)}, .., \hat{\mathsf{W}}^{(m)}, .., \hat{\mathsf{W}}^{(m')})$ *such that* $\|\mathbf{w}^{(m')}\|^2 + \ell((\hat{\mathsf{W}}^{(1)}, .., \hat{\mathsf{W}}^{(m)} .., \hat{\mathsf{W}}^{(m')}), \mathbf{w}^{(m')}) \leq \gamma$

Lemma 2 *Suppose* $\mathbb{E}[\|\hat{\mathbf{gp}}_t\|^2] \leq G^2$ *and* $\mathbb{E}[\|\hat{\mathbf{gq}}_t\|^2] \leq G^2$, *and we use the update rule of* $\mathsf{W}_{t+1} = \Pi_{\mathcal{W}}(\mathsf{W}_t - \eta_t(\hat{\mathbf{gp}}_t + \hat{\mathbf{gq}}_t))$ *where* $\mathbb{E}[\hat{\mathbf{gp}}_t] = \mathbf{gp}_t$ *and* $\mathbb{E}[\hat{\mathbf{gq}}_t] = \mathbf{gq}_t$. *If we use* $\eta_t = 1/(\lambda_1 + \lambda_2)t$, *then at time stamp* T

$$\mathbb{E}[\|\mathsf{W}_T - \mathsf{W}^\star\|^2] \leq \frac{12G^2}{(\lambda_1 + \lambda_2)^2 T}$$

Lemma 3 *Following the assumptions in lemma 2, but now we assume* $\eta_t = 1/t$ *since* λ_1 *and* λ_2 *are not always readily available, then started from* $\|\mathsf{W}_1 - \mathsf{W}^\star\|^2 \leq D$ *the convergence rate is bounded by*

$$\mathbb{E}[\|\mathsf{W}_T - \mathsf{W}^\star\|^2] \leq e^{-2\lambda(\ln T + 0.578)} D + (T-1)e^{-2\lambda \ln(T-1)} G^2$$

Based on lemma 1, 2,3 the author gives such theorem to prove that the proposed method will converge faster given same iteration times and learning steps.

Theorem 1 *Let* $\mathcal{P}(\mathsf{W})$ *be* λ_1*-strongly convex and* $\mathcal{Q}(\mathsf{W})$ *be* λ_2*-strongly convex near optimal* $\mathsf{W}^\star$ *and denote* $\mathsf{W}_T^{(F)}$ *and* $\mathsf{W}_T^{(\mathcal{P})}$ *as the solution after* T *iterations when applying SGD on* $F(\mathsf{W})$ *and* $\mathcal{P}(\mathsf{W})$ *respectively. Then our deeply supervised framework in eqn. (3) improves the the speed over using top layer only by* $\frac{\mathbb{E}[\|\mathsf{W}_T^{(\mathcal{P})} - \mathsf{W}^\star\|^2]}{\mathbb{E}[\|\mathsf{W}_T^{(F)} - \mathsf{W}^\star\|^2]} = \Theta(1 + \frac{\lambda_2^2}{\lambda_1^2}),\ when\ \eta_t = 1/\lambda t, \quad and,\ \frac{\mathbb{E}[\|\mathsf{W}_T^{(\mathcal{P})} - \mathsf{W}^\star\|^2]}{\mathbb{E}[\|\mathsf{W}_T^{(F)} - \mathsf{W}^\star\|^2]} = \Theta(e^{\ln(T)\lambda_2}),\ when\ \eta_t = 1/t.$

Constraints in hidden layer also help to learn more discriminative features.

Experimental results

Four benchmark datasets are used to evaluate the proposed method.

For MNIST dataset, the proposed method displays classification error of 0.39% under a single model without data whitening and augmentation and shows more generalized than traditional CNNs.

For CIFAR-10 DSN reaches classification error of 9.78% and 8.22% without and with data augmentation respectively. For CIFAR-100 DSN reaches 34.57% classification error.

For Street View House Numbers dataset, DSN reaches a classification error of 1.92%. All of the results listed above are better than previous technique such as stochastic pooling, maxout networks and so on.

Conclusion

This paper presents a deeply-supervised nets which enforces the traditional convolution neural network to learn discriminative feature maps and can reach optimal solution faster.

Summaries for "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation"

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author proposes a hybrid architecture combining deep Convolutional Neural Network and a Markov Random Field. This model can exploit structural domain constraints such as geometric relationships between body joint locations. The author claims their model outperforms existing state-of-the-art techniques.

Main idea

For the part detector, the author first denies the sliding-window ConvNet architecture which is used to slide over the input image to produce a dense hear-map output for each body-joint. The model is translation invariant. So the author combines an efficient sliding window-based architecture with multi-resolution and overlapping receptive fields. This model can be described as the following picture

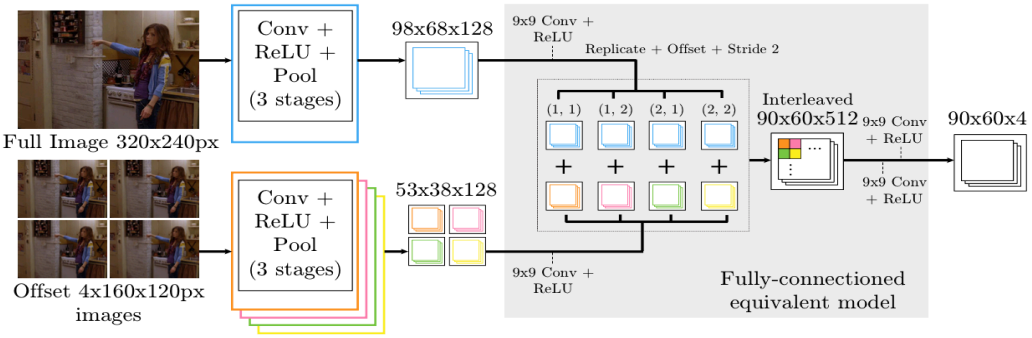


Figure 1 Efficient sliding window model with overlapping receptive fields

To improve training time the writers simplify the above architecture by replacing the lower-resolution stage with a single convolution bank shown as follows.

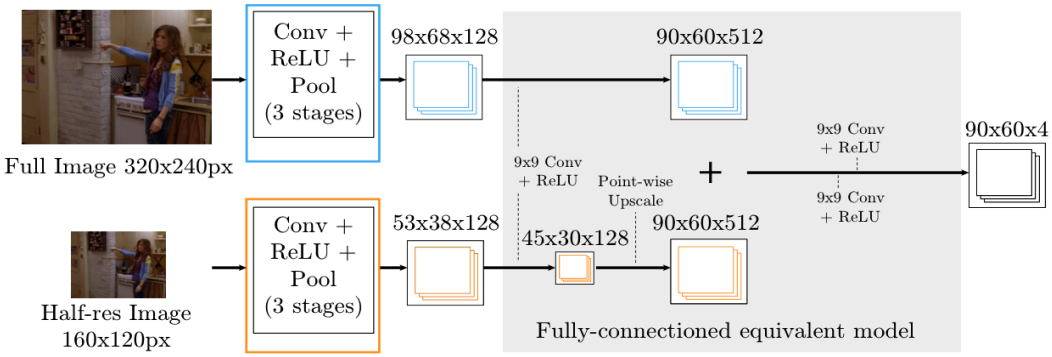


Figure 2 Approximation of the model in Figure 1

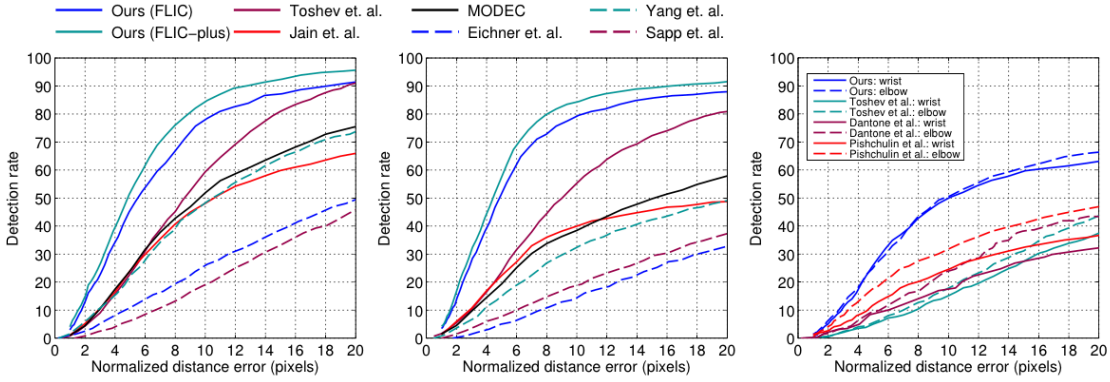
However, the author claims that the part-detector performance on the validation set predicts heat-maps that contain many false positive and poses that are anatomically incorrect. Furthermore, the author continues to formulate the spatial-model as an MRF-like model over the distribution of spatial locations for each body part. This model starts by connecting every body part to itself and to every other body part in a pair-wise fashion in the spatial model to create a fully connected graph. The final marginal likelihood $\overline{pA}$ can be described as

$$\overline{pA} = \frac{1}{Z}\prod_{v \in V}(pA|v * p_v + b_{v \to A})$$

For practical implementation the author treats the distributions as energies to avoid the evaluation of partition term of Z.

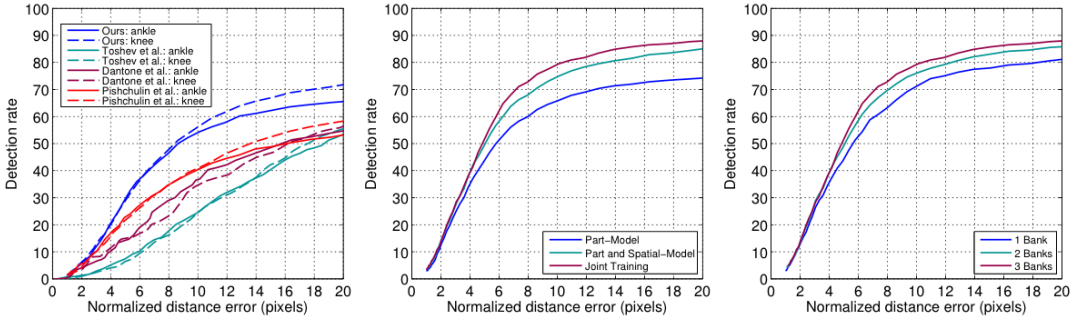
Experimental results

Two datasets are used to evaluate the proposed model: FLIC and extended-LSP. The results can be described as Figure 3 and Figure 4 shows.



(a) FLIC: Elbow (b) FLIC: Wrist (c) LSP: Wrist and Elbow

Figure 3 Model Performance



(a) LSP: Ankle and Knee (b) FLIC: Wrist (c) FLIC: Wrist

Figure 4 Model Performance

These datasets consist of still RGB images with 2D ground-truth joint information generated using Amazon Mechanical Turk. The author claims their model outperforms existing state-of the-art techniques on both of these challenging datasets with a considerable margin. Particularly, for large radii the model increases performance by 8 to 12%. Unified training of both models adds an additional 4-5% detection rate for large radii thresholds.

Conclusion

The author have shown that the unification of a novel ConvNet Part-Detector combing an MRF inspired by spatial model can significantly outperforms existing architecture on the task of human body pose recognition.