Gaussian Conditional Random Field Network for Semantic Segmentation

Raviteja Vemulapalli, Rama Chellappa
University of Maryland, College Park

Oncel Tuzel, Ming-Yu Liu
Mitsubishi Electric Research Laboratories, Cambridge

Main contribution: An end-to-end trainable deep network architecture that combines CNNs with a Gaussian conditional random field model.

Motivation

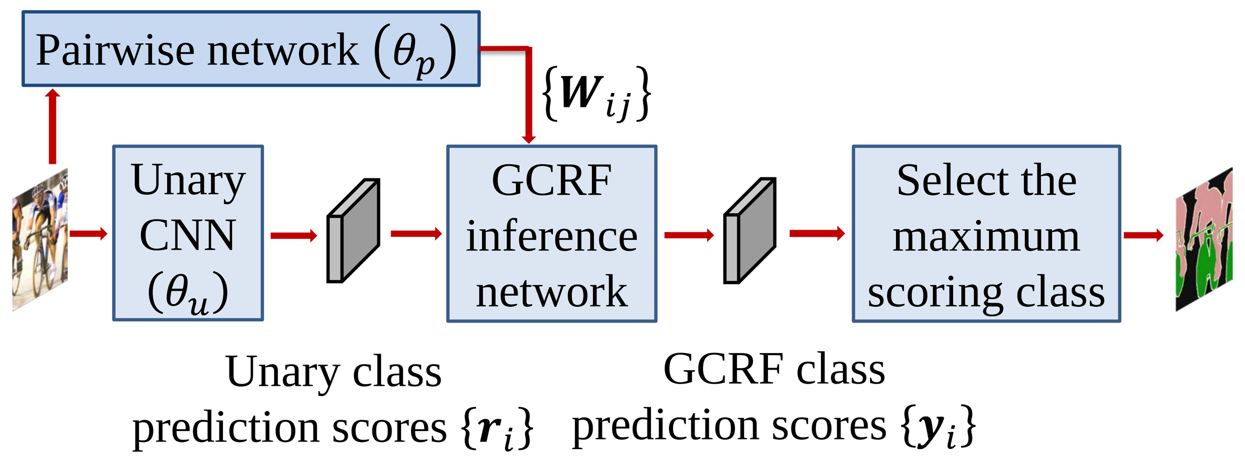
- CNNs do not explicitly model the interactions between output variables which is very important for structured prediction tasks such as semantic segmentation.
- Various recent approaches [1,2] combine CNNs with discrete CRFs.
- Inference techniques in the case of discrete CRFs do not have optimality guarantees upon convergence.
- In contrast, Gaussian mean field inference gives optimal solution upon convergence in the case of a Gaussian CRF.

Gaussian CRF Network

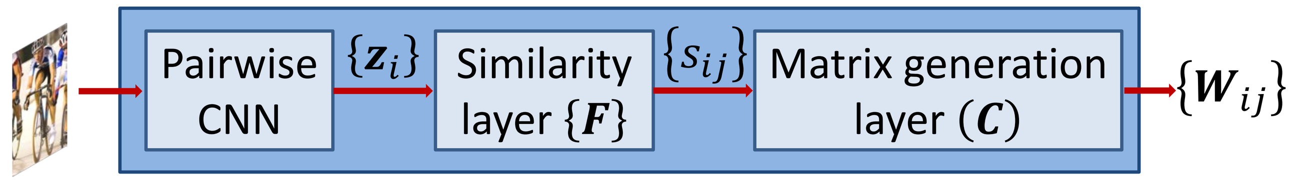
- We use a Gaussian CRF model on top of a CNN.
- Each discrete output variable is replaced by a vector of K continuous variables: $\boldsymbol{y}_i = [y_{i1}, \dots, y_{iK}] \in R^K$.
- y_{ik} represents the score for k^{th} class at i^{th} pixel.
- Class label for i^{th} pixel is given by $\underset{k}{\operatorname{argmax}}\, y_{ik}$.
- $\boldsymbol{X}$: Input image
- $P(\boldsymbol{Y}|\boldsymbol{X}) \propto e^{-\frac{1}{2}E}$, where

$$E = \underbrace{\sum_i \|\boldsymbol{y}_i - \boldsymbol{r}_i(\boldsymbol{X};\theta_u)\|_2^2}_{\text{Unary term}} + \underbrace{\sum_{ij} (\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boxed{\boldsymbol{W}_{ij}(\boldsymbol{X};\theta_p)} (\boldsymbol{y}_i - \boldsymbol{y}_j)}_{\text{Pairwise term}}; \quad \boldsymbol{W}_{ij} \succcurlyeq 0.$$

Pairwise potential parameters



Pairwise Network



- Each $\boldsymbol{W}_{ij}$ is computed as $\boldsymbol{W}_{ij} = s_{ij}\boldsymbol{C}; \boldsymbol{C} \succcurlyeq 0$.
- $s_{ij} \in [0,1]$ is a similarity measure between pixels i and j.
- $\boldsymbol{C}$ is a parameter matrix that encodes the class compatibility information.
- The similarity measure s_{ij} is computed as

$$s_{ij} = e^{-(\boldsymbol{z}_i - \boldsymbol{z}_j)^T \boldsymbol{F} (\boldsymbol{z}_i - \boldsymbol{z}_j)}.$$

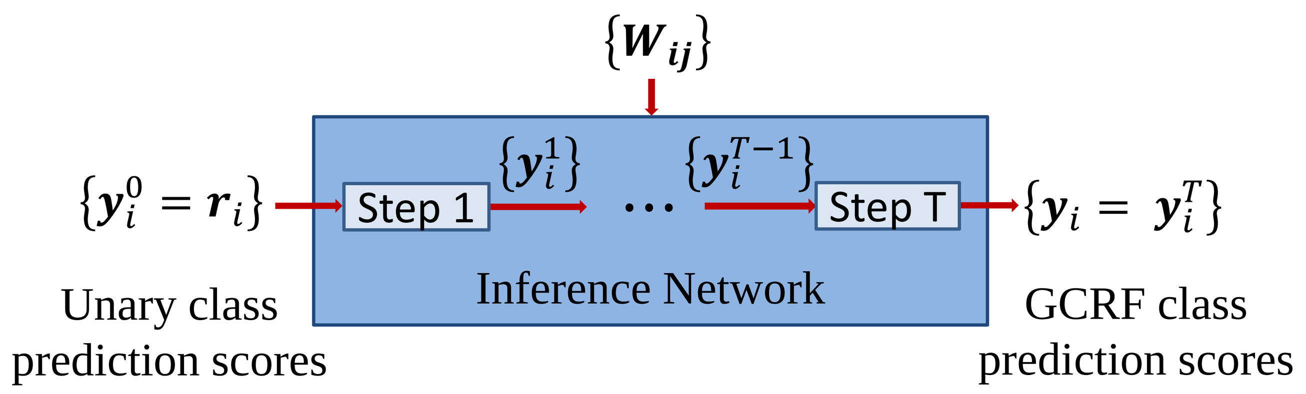
- $\boldsymbol{z}_i$ is a feature vector extracted at pixel i using the pairwise CNN.
- $\boldsymbol{F} \succcurlyeq 0$ is a parameter matrix that defines a Mahalanobis distance function.

Gaussian CRF Inference

$$\min_{\{\boldsymbol{y}_i\}} \sum_i \|\boldsymbol{y}_i - \boldsymbol{r}_i\|_2^2 + \sum_{ij} (\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{W}_{ij} (\boldsymbol{y}_i - \boldsymbol{y}_j)$$

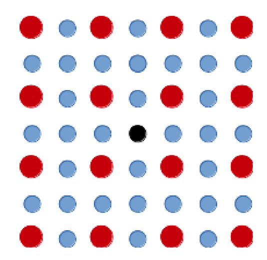
- We use the iterative Gaussian mean field (GMF) inference approach.
- We unroll the GMF inference steps into a deep network.
- GMF update equation (optimal coordinate descent step):

$$\boldsymbol{y}_i^{t+1} = \left(I + \sum_j \boldsymbol{W}_{ij}\right)^{-1} \left(\boldsymbol{r}_i + \sum_j \boldsymbol{W}_{ij} \boldsymbol{y}_j^t\right).$$



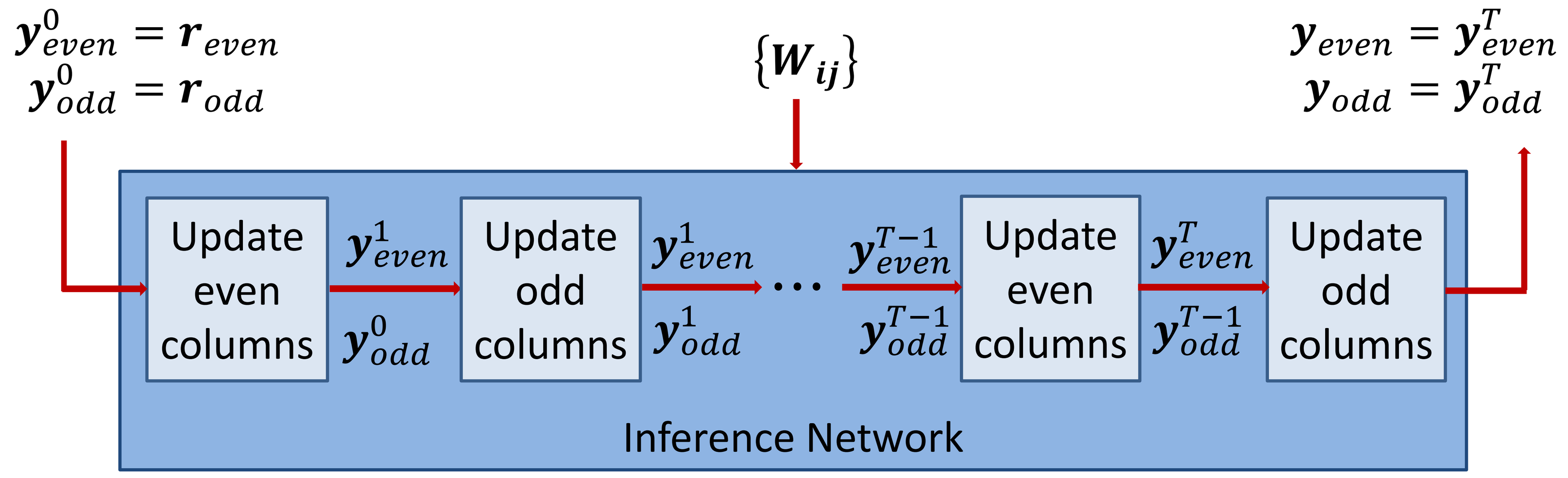
Convergence of GMF Inference

- For convergence, parallel GMF inference requires the precision matrix of the Gaussian distribution $P(\boldsymbol{Y}|\boldsymbol{X})$ to be diagonally dominant.
- In our graph, instead of connecting each pixel to every pixel, we connect each pixel to every other pixel along both rows and columns within a spatial neighborhood.



The center black pixel is connected only to the red pixels.

- This connectivity makes our graph bipartite.
  Even columns – Odd columns
- We can update all the pixels within a partition in parallel and still have convergence guarantees without the diagonal dominance constraints.
- This is equivalent to (optimal) block coordinate descent.

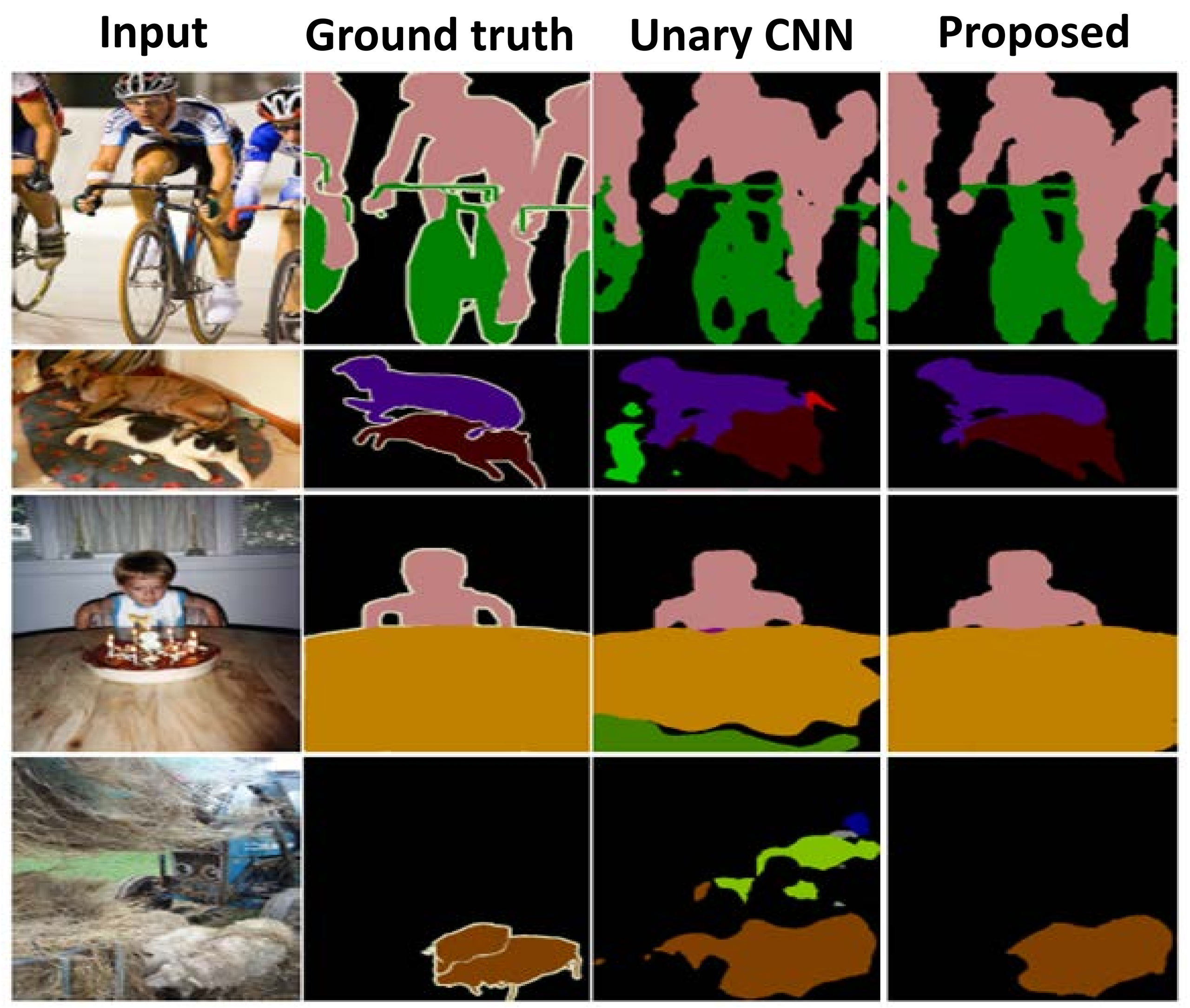


Experimental Results

- Initialized the CNNs using DeepLab CNN [1] pre-trained on ImageNet and finetuned on PASCAL VOC 2012.
- Pairwise CNN and similarity layer are pre-trained like a Siamese network at pixel level.
- GCRF network training loss function:

$$L(\{\boldsymbol{y}_i, l_i\}) = -\frac{1}{N}\sum_{i=1}^{N} \min\left(0,\ y_{il_i} - \max_{k \neq l_i} y_{ik} - T\right).$$

- The parameter matrices $\boldsymbol{C}$ and $\boldsymbol{F}$ are parameterized as $\boldsymbol{C} = \boldsymbol{R}\boldsymbol{R}^T$ and $\boldsymbol{F} = \sum_{m=1}^{M} f_m f_m^T$.
- We learn all the parameters (unary and pairwise CNN parameters, $\boldsymbol{R}$, $\{f_m\}$) by training the network end-to-end.



Results on PASCAL VOC 2012 test set when trained using ImageNet (for pretraining) and PASCAL VOC data.

Approaches that use CNNs and discrete CRFs																					
DeconvNet + CRF	92.9	87.8	41.9	80.6	63.9	67.3	88.1	78.4	81.3	25.9	73.7	61.2	72.0	77.0	79.9	78.7	59.5	78.3	55.0	75.2	61.5	70.5
object clique potentials	92.8	80.0	53.8	80.8	62.5	64.7	87.0	78.5	83.0	29.0	82.0	60.3	76.3	78.4	83.0	79.8	57.0	80.0	53.1	70.1	63.1	71.2
DeepLab CNN-CRF	93.3	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN	94.0	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet + FCN + CRF	93.1	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
Proposed GCRF network	93.4	85.2	43.9	83.3	65.2	68.3	89.0	82.7	85.3	31.1	79.5	63.3	80.5	79.3	85.5	81.0	60.5	85.5	52.0	77.3	65.1	73.2

References:

1. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In ICLR, 2015.
2. S. Zheng, S. Jayasumana, B. R.-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In ICCV, 2015.