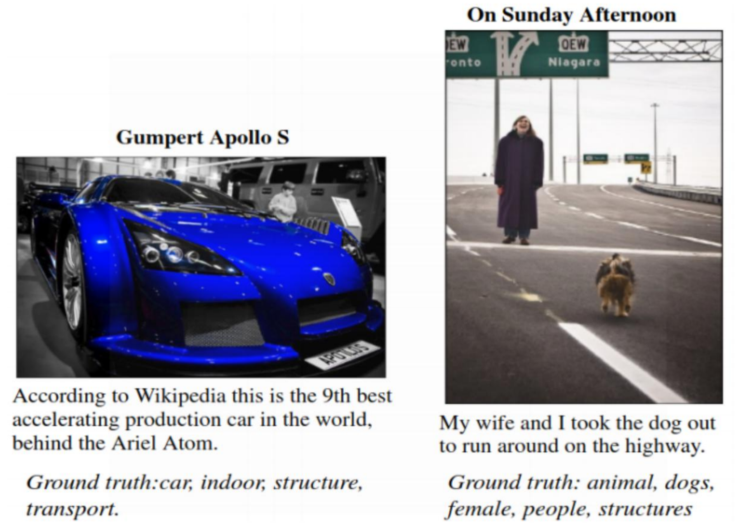
Deep Neural Networks In Fully Connected CRF For Image Labeling With Social Network Metadata

Chengjiang Long Roddy Collins Eran Swears Anthony Hoogs
Kitware Inc. (1712 Route 9 Suite 300, Clifton Park, NY 12065)
{chengjiang.long, roddy.collins, eran.swears, anthony.hoogs}@kitware.com

IEEE 2019 Winter Conference on Applications of Computer Vision

Introduction

Observation: Social multimedia dataset contains (1) images, (2) text information like title, description, comments, and (3) other meta information like user information, image gallery, uploder-defined groups, and links between shared contents.



Intuition: We hypothesize that using social media context jointly with pixel information should improve the state-of-the-art in image labeling

Goal: We seek to understand the relative contribution of pixels, text and other information in predicting image labels.

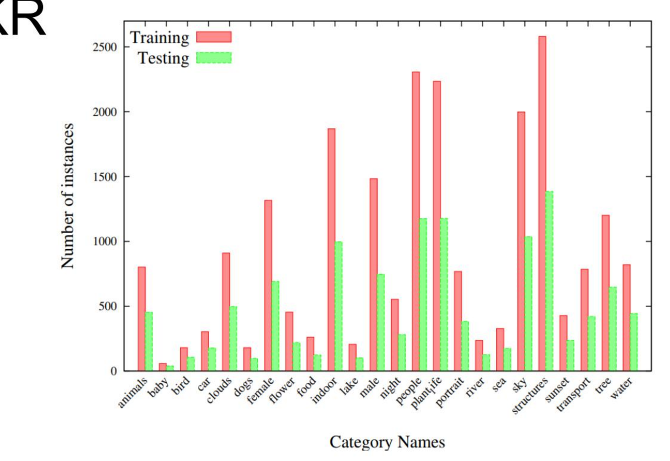
Competing Algorithms

[McAuley-CRF] J. J. McAuley and J. Leskovec. Image labeling on a network:Using social-network metadata for image classification. In ECCV, 2012.

[Jhon-NCNN] J. Johnson et al. Love thy neighbors: Image annotation by exploiting image metadata. In ICCV 2015.

MIR-9K Dataset

A subset of the MIRFLICKR dataset, contains 6000 + 3182 images with 24 categories.
It involves a set of 3,213 users, a collection of 34,942 words and 17,687 image groups.



Proposed Approach

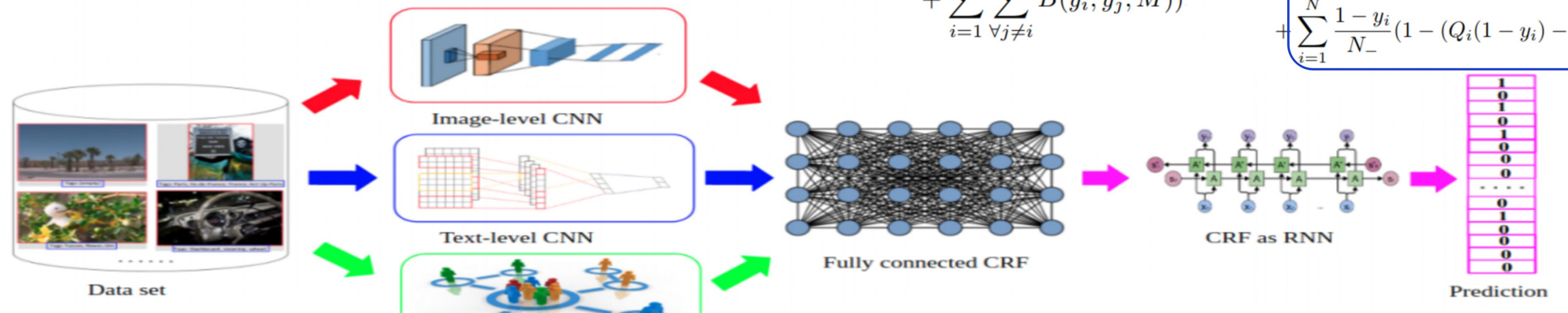
Joint probability:

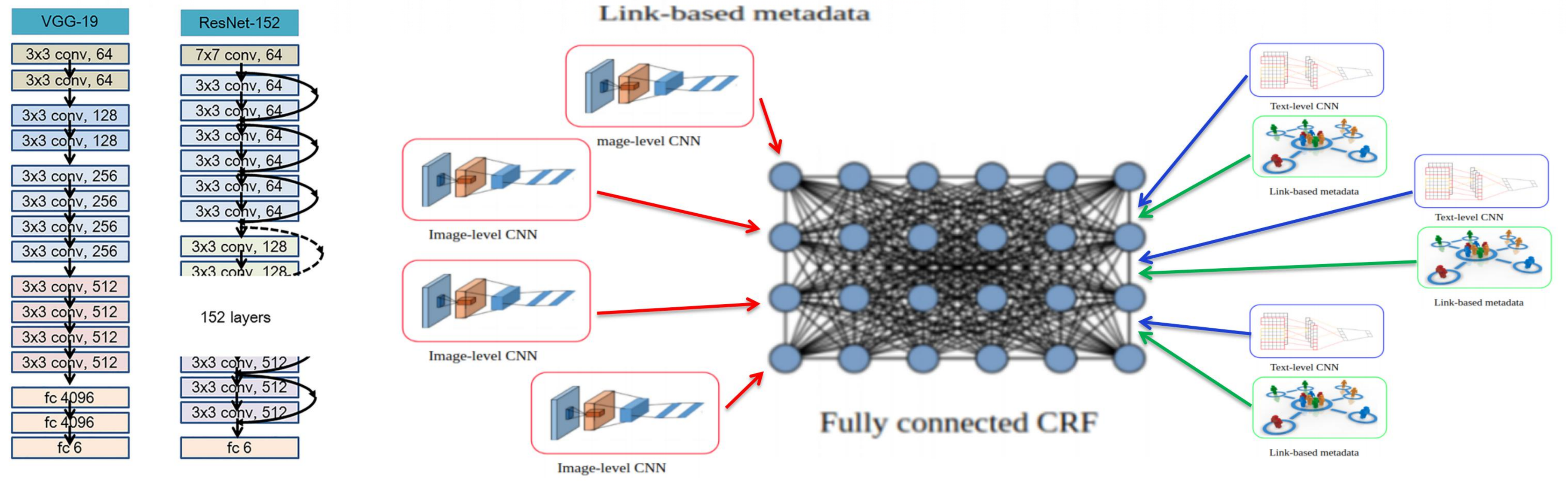
$$P(Y|I,M) = P(Y|\mathbf{x},M) = \frac{1}{Z}\exp\Big(\sum_{i=1}^{N} A(y_i,\mathbf{x}_i) + \sum_{i=1}^{N}\sum_{\forall j\neq i} B(y_i,y_j,M)\Big)$$

$$Loss(Q,Y) = \underbrace{\sum_{i=1}^{N} -\frac{y_i}{N_+}\log Q_i(y_i) - \frac{1-y_i}{N_-}\log Q_i(1-y_i)}_{\text{BCE}} + \underbrace{\sum_{i=1}^{N}\frac{y_i}{N_+}(1-(Q_i(y_i)-Q_i(1-y_i))) + \sum_{i=1}^{N}\frac{1-y_i}{N_-}(1-(Q_i(1-y_i)-Q_i(y_i)))}_{\text{Ranking loss}} + \lambda\|\mathbf{w}_A\|_2,$$

BCE

Ranking loss

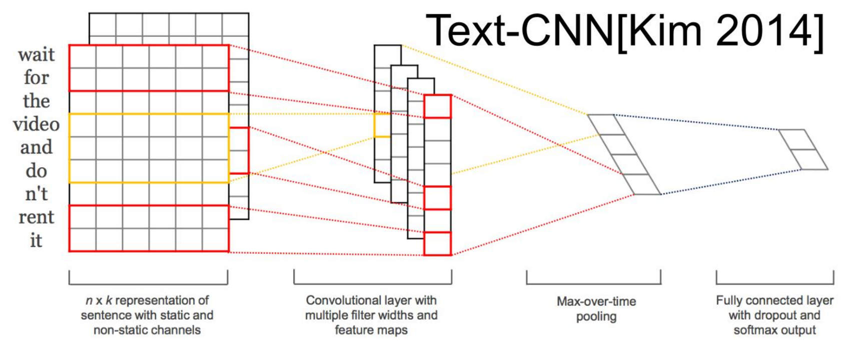




$$A(y_i,\mathbf{x}_i) = \mathbf{w}_A^{y_i}\mathbf{x}_i + \mathbf{b}_A^{y_i}$$

$$B(y_i,y_j,M)) = \mu(y_i,y_j)\sum_{k\in\{text,set,group\}}\mathbf{w}_B^k S_k(i,j)$$

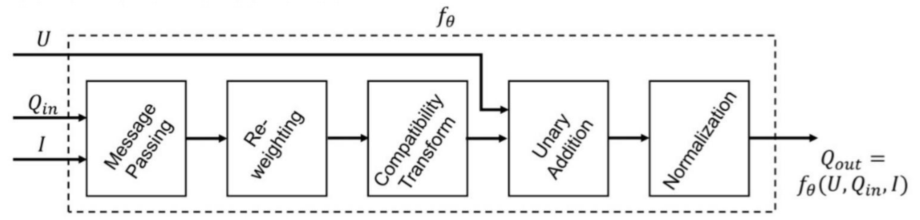
Text-CNN[Kim 2014]



$$S_{text}(i,j) = \exp\Big(-\frac{|\mathbf{x}_i^{text}-\mathbf{x}_j^{text}|^2}{2\theta_{text}}\Big)$$

$$\left.\begin{aligned} S_{set}(i,j) &= \exp\Big(-\frac{d(i,j,M_{set})^2}{2\theta_{set}}\Big)\\ S_{group}(i,j) &= \exp\Big(-\frac{d(i,j,M_{group})^2}{2\theta_{group}}\Big)\end{aligned}\right\}$$

d is based on Jaccard similarity between two nodes.



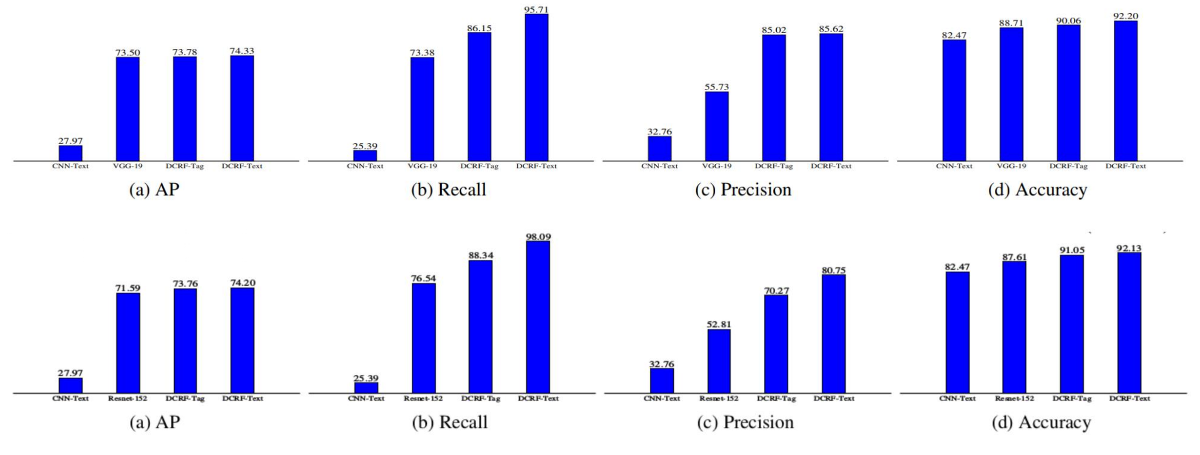
Algorithm 1: The outline of our proposed DCRF algorithm

Input: I and M
Output: Q

$\mathbf{x} \leftarrow \mathrm{CNN}_{image}(I)$
$\mathbf{x}^{text} \leftarrow \mathrm{CNN}_{text}(M)$
$\mathbf{t}^{set}, \mathbf{t}^{group} \leftarrow M$
$U \leftarrow \mathbf{w}_A\mathbf{x} + \mathbf{b}_A$
$Q_i(y) \leftarrow \frac{1}{Z_i}\exp\{U_i(y)\}$
while *not converged* **do**
  $\tilde{Q}_i^{(k)}(y) \leftarrow \sum_{\forall j\neq i} S_k(i,j)Q_j(y)$ for all k
  $\check{Q}_i(y) \leftarrow \sum_k \mathbf{w}_B^k\tilde{Q}_i^{(k)}(y)$
  $\hat{Q}_i(y) \leftarrow \sum_{y'}\mu(y,y')\check{Q}_i(y)$
  $\breve{Q}_i(y) \leftarrow U_i(y) - \hat{Q}_i(y)$
  $Q_i(y) \leftarrow \frac{1}{Z_i}\exp\{\breve{Q}_i(y)\}$
end

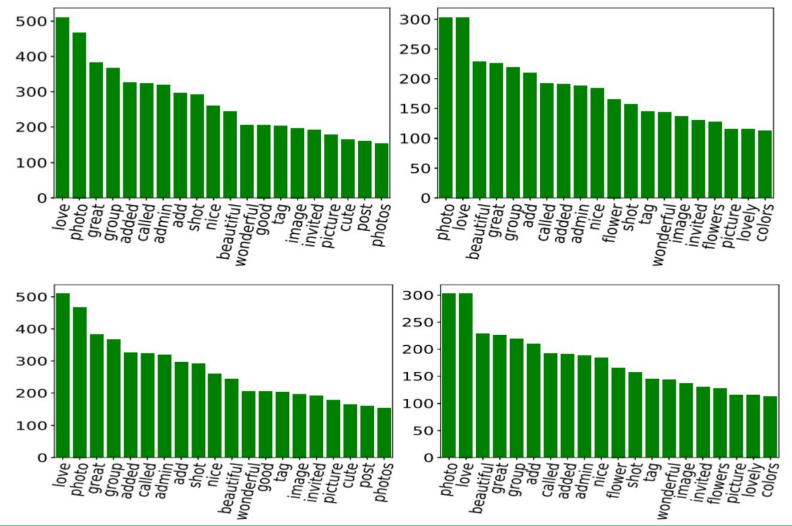
Experiments

Effectiveness of the text-level CNN



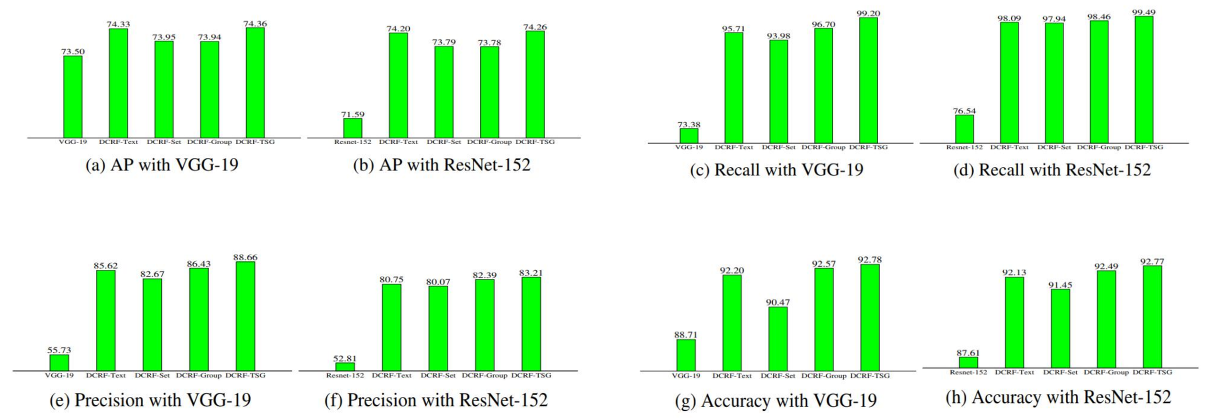
With VGG-19 Network

With ResNet-152 Network



Animal	Flower
Portrait	Water

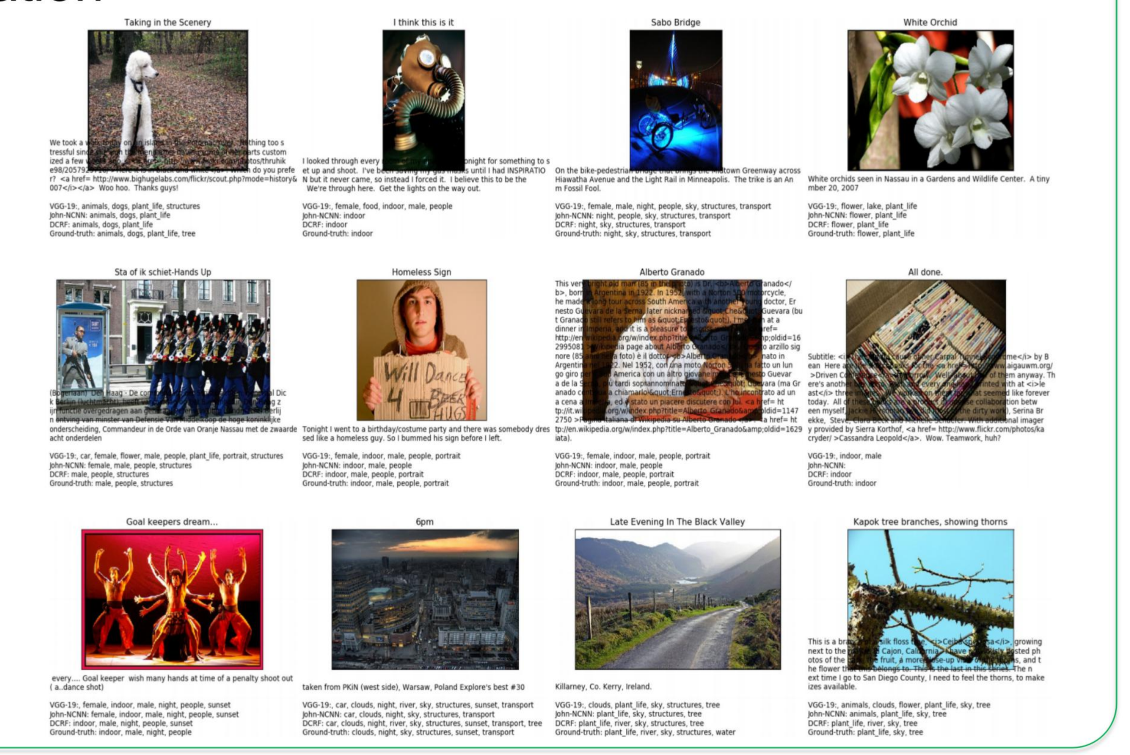
Effectiveness of metatdata



Compare with the state-of-the-art approches

	AP	REC	PRE	ACC
CNN_{text} [15]	27.97	25.39	32.76	82.47
$\mathrm{AlexNet}_{img}$ [17]	62.54	76.30	40.25	74.56
$\mathrm{VGG\text{-}19}_{img}$ [27]	**73.50**	**77.38**	**55.73**	**88.71**
$\mathrm{ResNet\text{-}152}_{img}$ [9]	71.59	76.54	52.82	87.62
$\mathrm{DenseNet\text{-}201}_{img}$ [10]	63.26	72.55	42.93	85.06
McAuley-CRF [21]	54.73	40.75	59.44	83.1
$\mathrm{John\text{-}NCNN}_{vgg}$ [12]	**73.78**	**61.18**	79.01	**92.57**
$\mathrm{John\text{-}NCNN}_{res}$ [12]	72.90	50.59	**81.39**	91.87
DCRF_{vgg}-BCE	74.13	92.66	85.86	92.50
DCRF_{vgg}-RLoss	74.29	93.12	88.18	92.61
DCRF_{vgg}-BCE+RLoss	**74.36**	99.20	**88.66**	**92.78**
DCRF_{res}-BCE	74.05	91.52	74.69	91.74
DCRF_{res}-RLoss	74.09	94.38	77.59	91.93
DCRF_{res}-BCE+RLoss	74.26	**99.49**	83.21	92.77

Visualization



Conclusion and future work

We propose a novel deep fully connected CRF based framework with a joint end-to-end CNN-RNN formulation for image labeling which combines the strengths of both CNNs and RNNs.

Our future work includes investigating more effective meta information, and improving the efficiency of the current DCRF framework to handle more complicated real-world application problems.