Blending the Past and Present of Automatic Image Annotation

Ayushi Dutta

Roll: 201507668

Advisors: ***Prof. C. V. Jawahar*** **(IIIT-Hyderabad),**

Assistant Prof. Yashaswi Verma **(IIT-Jodhpur)**

Center for Visual Information Technology,

IIIT-Hyderabad, India

Automatic Image Annotation

nighttime, person, running, street, vehicle

- An image is tagged with a set of discrete labels from a given vocabulary that best describe the image
- The labels may refer to different objects/things , scene/context, concepts, actions etc..

Challenges

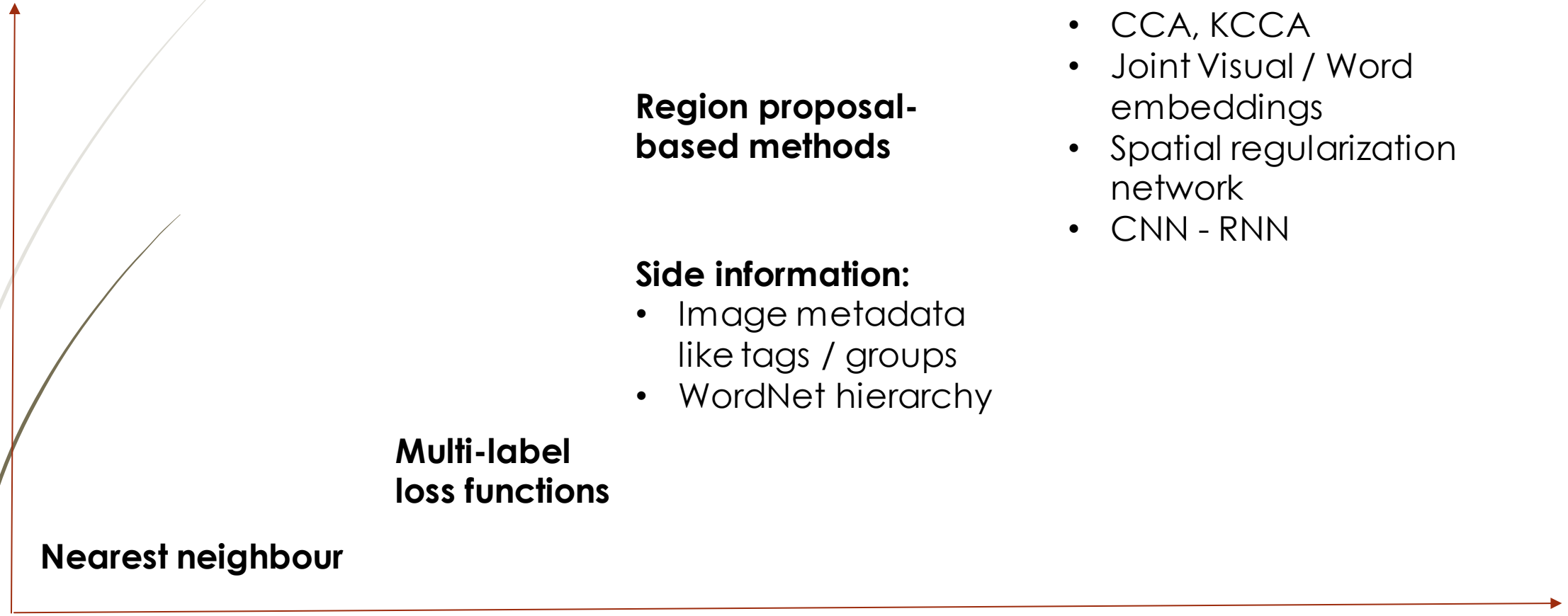
nighttime, person, running, street, vehicle

- Single label to multi labels
- Variable no. of labels
- Scale of objects present in the image
- Occlusions or unconventional views
- Image Quality
- Abstract representations
- Weakly supervised problem
- Dataset related challenges: class imbalance, incomplete labelling, label ambiguity

Applications

- Image understanding, real world images
- Archiving and accessing large image collections
- Aid search engines in image retrieval
- Other visual learning tasks such as image captioning, scene recognition, multi-object recognition etc..

Previous Work



The Context

- As compared to single label image classification, improvement in multi label image annotation is gradual and quite challenging.
- There are fundamental issues in the dataset properties and performance evaluation metrics, that affect the performance of annotation methods of existing approaches.

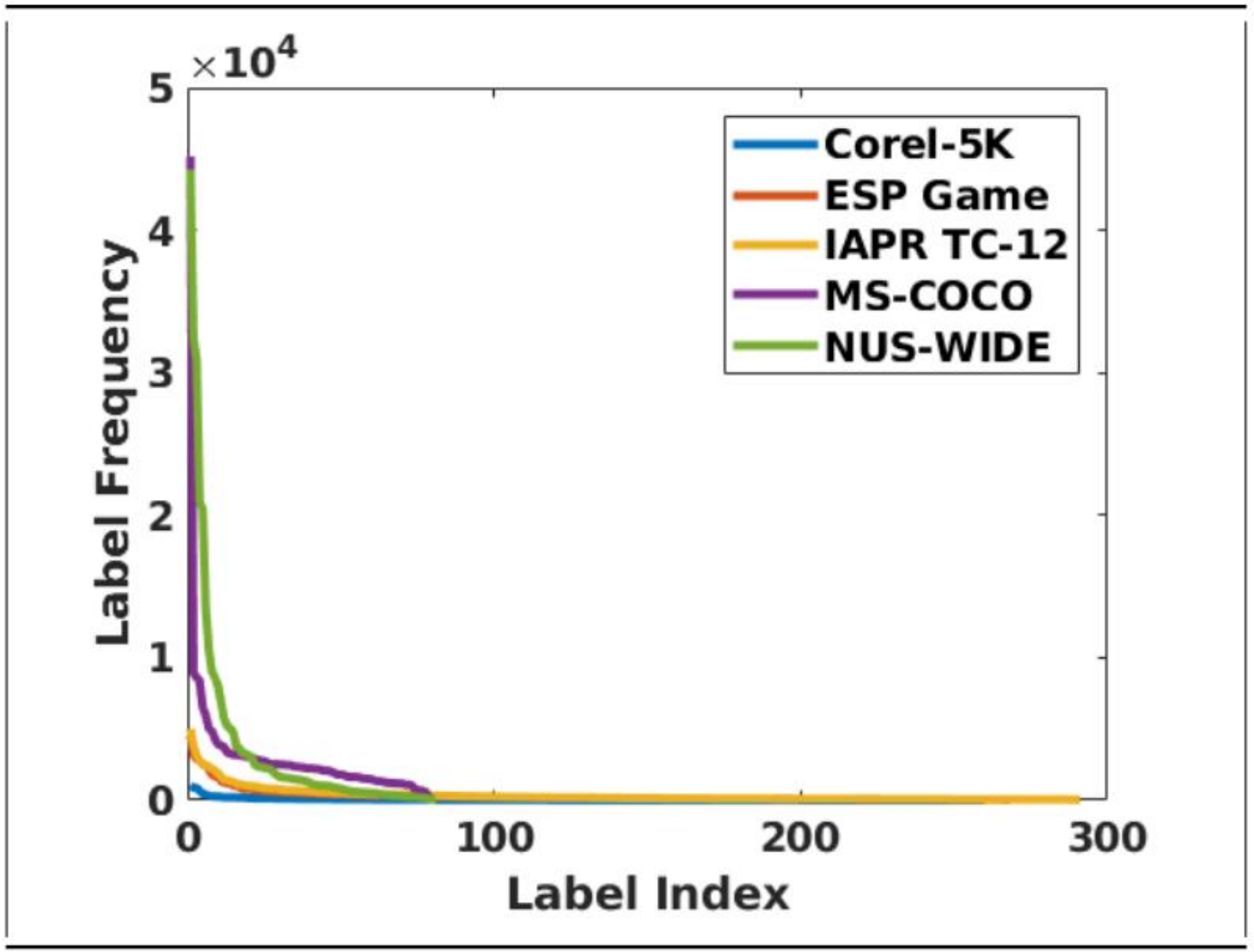
In this Thesis

- Empirically analyze the core issues/aspects in the image annotation domain
  - Relative trade-off of per-label versus per-image performance evaluation metrics
  - Quantify the degree of diversity (both image as well as label) in image annotation datasets
- Propose a new CNN-RNN-style model for image annotation
  - Overcome the limitation of earlier CNN-RNN models regarding training in a pre-defined label order.
  - Explore the scope and effectiveness of CNN-RNN-style models

Background

Datasets

Dataset	No. of training images	No. of test images	No. of labels	Avg no. of labels/image
Corel-5K	4500	499	260	3.4
ESP Game	18689	2081	268	4.7
IAPR TC-12	17665	1962	291	5.7
NUS-WIDE	125449	83898	81	2.4
MS-COCO	82783	40504	80	2.9



Frequency of labels in the training sets of each of the five datasets sorted in decreasing order

Performance Evaluation Metrics

Images	Label 1	Label 2	Label 3	Label 4	Label 5
1	1	0	1	0	1
2	0	0	1	0	1
3	1	0	1	1	0
4	0	1	0	1	0
5	0	0	1	0	1

Per Image

Per Label

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

$\text{N}+$ The number of labels with positive recall

Benchmark image annotation techniques

- **Non-Deep Learning**
  - **JEC:** *A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In ECCV, 2008*
  - **TagRel:** *X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. Trans. Multi., 11(7):1310–1322, 2009*
  - **TagProp:** *M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbour models for image auto-annotation. In ICCV, 2009*
  - **2PKNN:** *Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In ECCV, 2012*
  - **SVM:** *N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods. Cambridge University Press, 2000*
- **Deep Learning**
  - **Softmax Cross Entropy**
  - **Sigmoid Cross Entropy**
  - **Pairwise Ranking**
  - **WARP:** *Y. Gong, Y. Jia, T. K. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. In ICLR, 2014*
  - **LSEP:** *Y. Li, Y. Song, and J. Luo. Improving pairwise ranking for multi-label image classification. In CVPR, 2017*

CVIT

Non-deep learning based methods

Joint Equal Contribution (JEC)	Nearest neighbour based, labels chosen based on high frequencies and co-occurrence with the initially assigned labels
Tag Relevance (TagRel)	Nearest neighbour based, penalizes very high frequency based on overall frequency of that label
TagProp	Parametric weighted(distance) nearest neighbour based, maximize the likelihood of true labels during training
2PKNN	Weighted(distance) nearest neighbour based, balanced neighbourhood
Support Vector Machines (SVM)	A one-vs-rest binary classifier learnt for each label

Deep Learning based methods

- **Softmax Cross Entropy**

Ground-truth probability p_i by normalizing y_i as $y_i / ||y_i||_1$. Posterior probability $\hat{p}_{ij} = \frac{\exp(f_j(x_i))}{\sum_{k=1}^{c} \exp(f_k(x_i))}$

- **Sigmoid Cross Entropy** $\hat{p}_{ij} = 1/(1 + \exp(-f_j(x_i)))$.

- **Pairwise Ranking** $J = \sum_{i=1}^{n} \sum_{j=1}^{c+} \sum_{k=1}^{c-} \max(0, 1 - f_j(x_i) + f_k(x_i))$

- **Weighted Approximate Ranking(WARP)** $J = \sum_{i=1}^{n} \sum_{j=1}^{c+} \sum_{k=1}^{c-} L(r_j) \max(0, 1 - f_j(x_i) + f_k(x_i))$

- **Log-Sum-Exp Pairwise Ranking(LSEP)** $J = \sum_{i=1}^{n} \log(1 + \sum_{v \notin y_i} \sum_{u \in y_i} \exp(f_v(x_i) - f_u(x_i)))$

CVIT

Contribution-1

The Quirks of What Works

Experimental Setup

5 Datasets: Corel-5K, ESP Game, IAPR TC-12, NUS-WIDE, MS-COCO

Baseline CNN features (GoogLeNet and ResNet)

Perform 10 benchmark image annotation techniques:

- Non-deep learning – JEC, TagRel, TagProp, 2PKNN, SVM
- Deep Learning – Softmax CE, Sigmoid CE, Pairwise Ranking, WARP, LSEP

Based on scores, assign top k labels to each test image in the dataset.
K=5 for Corel, ESP, IAPR; K=3 for NUSWIDE, COCO

Empirical Analyses

Per-label versus Per-image Evaluation

- **In per-image metrics** each test image contributes equally, and thus they tend to get biased towards performance on frequent labels.
- **In per-label metrics** each label contributes equally, due to which they tend to get affected by performance on rare labels.

Model Comparison

Method	Per-label metrics					Per-image metrics			
	P_L	R_L	$F1_L$	mAP_L	N+	P_I	R_I	$F1_I$	mAP_I
SoftMax*	45.16	51.72	48.22	46.45	81	52.98	74.92	62.07	79.95
Sigmoid*	45.91	52.18	48.85	53.97	81	53.84	75.69	62.92	81.19
Ranking*	44.49	51.70	47.82	45.41	81	52.84	74.27	61.75	79.34
WARP*	43.91	53.17	48.09	46.04	81	53.03	74.56	61.98	79.54
LSEP*	44.29	53.46	48.45	49.32	81	53.64	75.54	62.73	80.91
JEC	37.15	40.91	38.94	21.63	80	29.36	69.32	41.25	61.68
TagRel	39.75	59.27	47.58	49.15	81	49.87	70.71	58.49	73.55
TagProp	48.84	58.10	53.07	53.81	80	51.52	73.16	60.46	76.90
2PKNN	52.49	52.28	52.38	51.96	81	45.30	64.77	53.31	67.82
SVM	46.56	52.38	49.30	51.84	81	53.31	74.96	62.31	79.87

Performance comparison of various annotation models (deep-learning based models are marked by '*') on the **NUS-WIDE** dataset using GoogLeNet features

Experiment with upper-bounds of various methods

- We assign a fixed number of top K labels to each test image during evaluation.
- We try to evaluate the best performance achievable by each method.
- Ordered list of labels(Rare to Freq): l1, l2, l3, ...l(n-1), l(n)

GT: animal, road, zebra

Method Prediction: animal, sky, zebra

Rare: animal, zebra, **l1**

Freq: animal, zebra, **l(n)**

Random: animal, zebra, **random(l1,.. ,l(n))**

	Method	Per-label F1 score ($F1_L$)				Per-image F1 score ($F1_I$)			
		True	Rare	Freq.	Rand	True	Rare	Freq.	Rand
NUS-WIDE	Ground	100.00	98.56	99.11	62.33	100.00	79.88	79.90	79.88
	SoftMax*	48.22	68.33	67.97	36.41	62.07	62.08	65.12	62.48
	Sigmoid*	48.85	68.57	68.30	36.49	62.92	62.94	65.36	63.33
	Ranking*	47.82	68.34	67.94	36.25	61.75	61.76	64.82	62.20
	WARP*	48.09	69.44	69.19	36.99	61.98	61.99	65.12	62.43
	LSEP*	48.45	69.58	69.40	37.17	62.73	62.75	65.51	63.16
	JEC	38.94	57.17	56.62	28.81	41.25	56.74	61.01	57.37
	TagRel	47.58	74.31	74.42	39.42	58.49	58.50	65.80	59.07
	TagProp	53.07	73.61	72.96	39.51	60.46	60.48	65.13	60.95
	2PKNN	52.38	68.03	68.64	33.88	53.31	53.32	58.65	54.06
	SVM	49.30	68.95	68.48	36.60	62.31	62.32	64.89	62.72

Comparing the actual per-label and per-image performance (using GoogLeNet) of various label prediction models (deep-learning based models are marked by *) with those obtained by **replacing incorrect predictions with rare/frequent/random incorrect labels.**

Experiment with rare / frequent / random label assignments

Dataset	Method	Per-label metrics				Per-image metrics		
		P_L	R_L	$F1_L$	N+	P_I	R_I	$F1_I$
NUS-WIDE	Rare	0.00	3.70	0.00	3	0.04	0.06	0.05
	Frequent	1.06	3.70	1.65	3	28.71	39.37	33.21
	Random	2.95	3.68	3.28	80	2.96	3.71	3.30

Performance by **assigning the three most rare, the three most frequent, and three randomly chosen labels** to each test image

Dataset Label Diversity

			
animal, grass, mountain, sky	animal, grass, mountain, sky	animal, grass, mountain, sky	animal, grass, mountain, sky

Example of **images belonging to the same label-set from NUS-WIDE dataset**, both training (first two from left) and test (last two from left) images

- Label-set is the set of labels assigned to an image
- We define two measures: percentage unique label-sets and novel label-sets.

$$\text{unique}_{\text{test}} = \frac{|\text{unique label-sets}_{\text{test}}|}{|\text{test images}|} * 100$$

$$\text{novel}_{\text{test}} = \frac{|\text{label-sets}_{\text{test}} - \text{label-sets}_{\text{train}}|}{|\text{test images}|} * 100$$

Results: Unique and Novel label-sets in test datasets

Dataset	Corel-5K	ESP Game	IAPR TC-12	NUS-WIDE	MS-COCO
Unique	86.8	95.2	95.3	12.3	27.2
Novel	48.9	82.5	73.2	6.7	17.3

Dataset Image Diversity

- Test set should contain compositionally novel images that are not seen in entirety in the training set
- We identify compositionally novel as well as compositionally similar images in the test sets and evaluate the performance of different methods on these subsets.

Proposed Approach

For each test image, compute the mean distance of 50 closest images from the training images

Pick the 20% most overlapping and 20% least overlapping images

Compute the performance of each method on the two subsets.



Examples from the **"most" overlapping images** from the NUS-WIDE dataset



Examples from the **"least" overlapping images** from the NUS-WIDE dataset

Performance on most overlapping images

	Method	Per-label metrics				Per-image metrics			
		P_L	R_L	$F1_L$	mAP_L	P_I	R_I	$F1_I$	mAP_I
NUS-WIDE	SoftMax*	59.80	65.98	62.74	63.53	61.75	81.68	70.33	87.64
	Sigmoid*	60.18	65.76	62.84	69.02	62.48	82.33	71.05	88.93
	Ranking*	57.75	66.30	61.73	62.91	61.79	81.54	70.30	87.64
	WARP*	58.08	67.26	62.34	62.94	61.94	81.69	70.46	87.48
	LSEP*	59.23	66.77	62.77	65.60	62.51	82.38	71.08	88.86
	JEC	56.54	49.29	52.67	38.02	56.55	75.62	64.71	71.22
	TagRel	58.44	70.99	64.12	66.16	59.69	79.20	68.08	82.76
	TagProp	65.64	71.83	68.59	68.93	60.72	80.44	69.20	85.18
	2PKNN	65.26	69.97	67.53	66.94	58.74	77.97	67.00	81.63
	SVM	63.90	67.79	65.79	66.67	61.57	81.16	70.02	86.95

Performance comparison (using GoogLeNet) of various label prediction models (deep-learning based models are marked by *) over the **20% most(top) overlapping test subsets**

Performance on least overlapping images

	Method	Per-label metrics				Per-image metrics			
		P_L	R_L	$F1_L$	mAP_L	P_I	R_I	$F1_I$	mAP_I
NUS-WIDE	SoftMax*	23.32	25.14	24.19	18.86	42.30	67.14	51.90	67.45
	Sigmoid*	23.02	26.17	24.49	22.49	42.86	67.72	52.49	68.26
	Ranking*	20.40	24.06	22.08	16.86	41.31	64.89	50.48	65.15
	WARP*	20.94	25.80	23.11	17.74	41.59	65.53	50.89	65.76
	LSEP*	21.95	27.31	24.34	19.79	42.68	67.53	52.30	67.88
	JEC	16.91	24.06	19.86	10.10	37.29	58.77	45.63	48.18
	TagRel	17.53	35.18	23.40	20.07	38.12	60.38	46.74	58.92
	TagProp	25.17	27.97	26.50	22.78	41.72	66.01	51.12	65.02
	2PKNN	29.08	19.87	23.61	21.02	34.02	53.66	41.64	50.03
	SVM	21.41	24.16	22.70	20.90	42.53	66.96	52.02	67.17

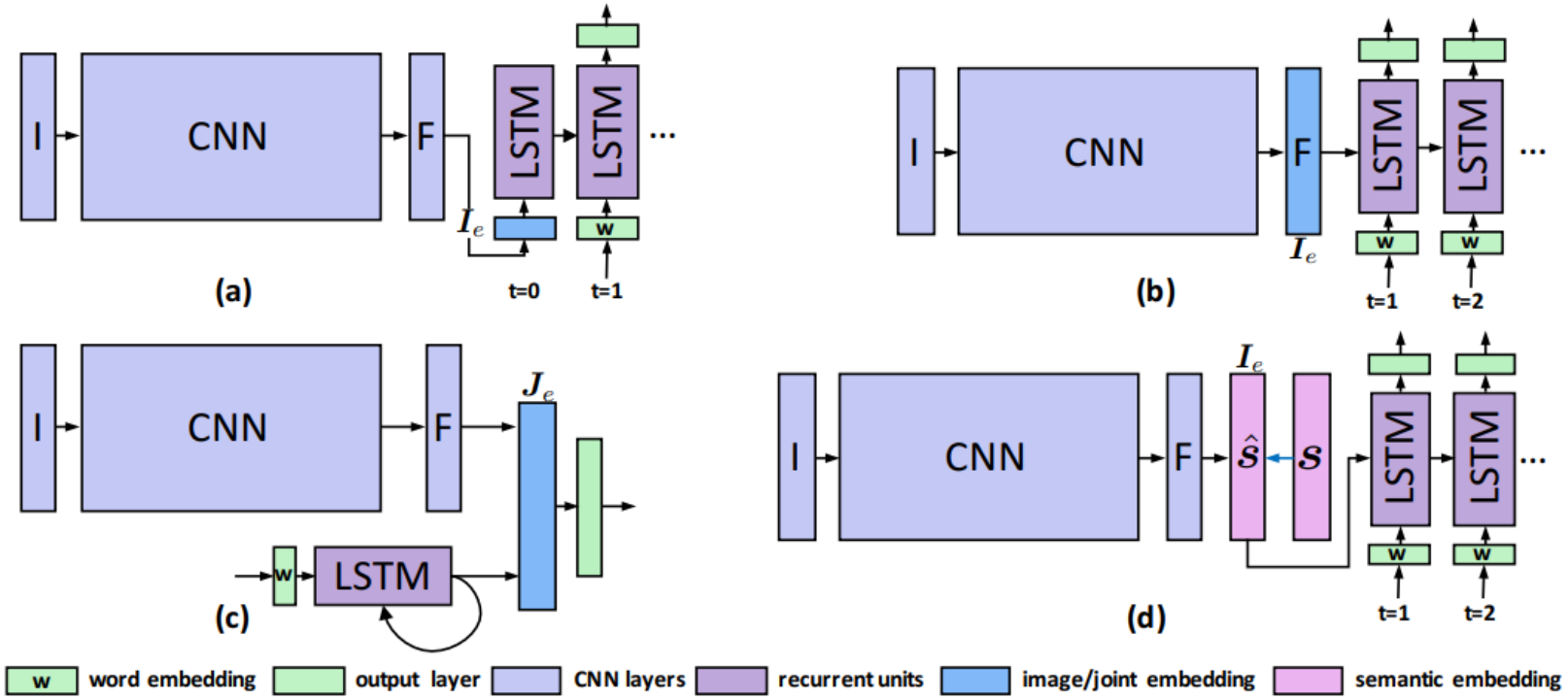
Performance comparison (using GoogLeNet) of various label prediction models (deep-learning based models are marked by *) over the **20% least(top) overlapping test subsets**

Conclusion

- Per-label metrics should be preferred over per-image metrics for comparing image annotation techniques in general.
- Datasets should be designed keeping in mind the image and label diversity. This also reduces the impact of rare and frequent labels on the performance of various annotation methods.

Contribution-2

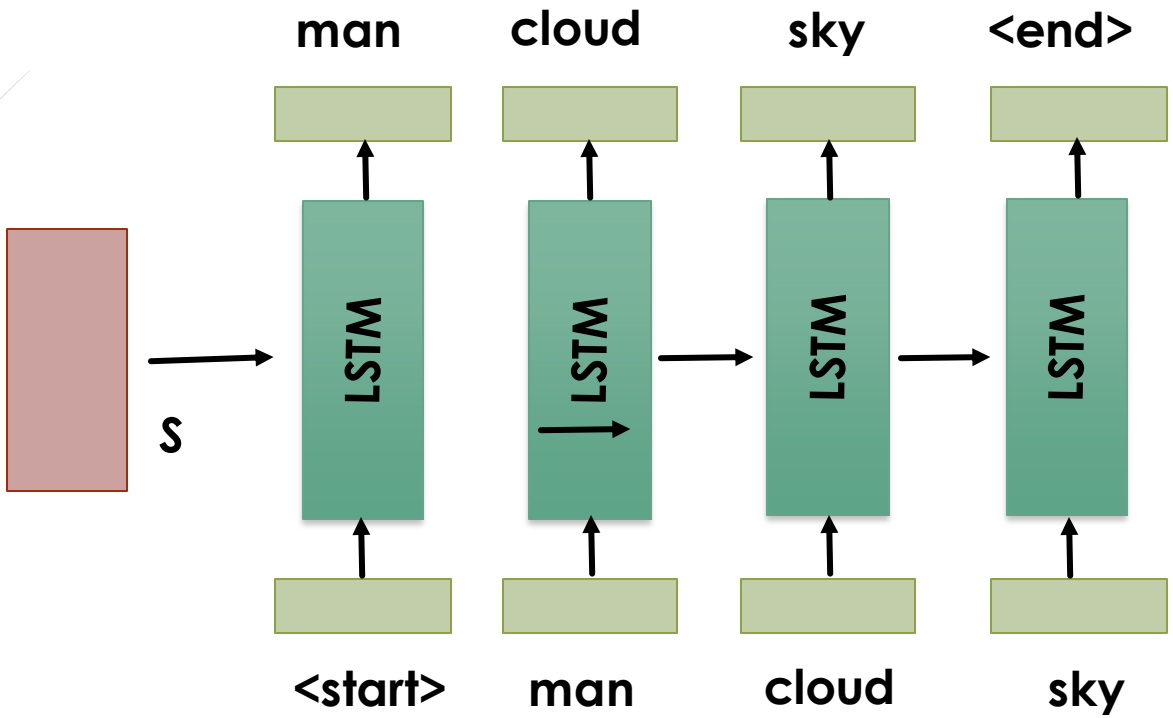
Revisiting CNN-RNN



F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. Semantic regularisation for recurrent image annotation. In CVPR, 2017

- Multiple labels of an image share rich semantic relationships among them
- Motivation to create a deep unified framework
- RNN's capacity to capture higher-order label relationships

Sequential Semantic Regularized CNN-RNN

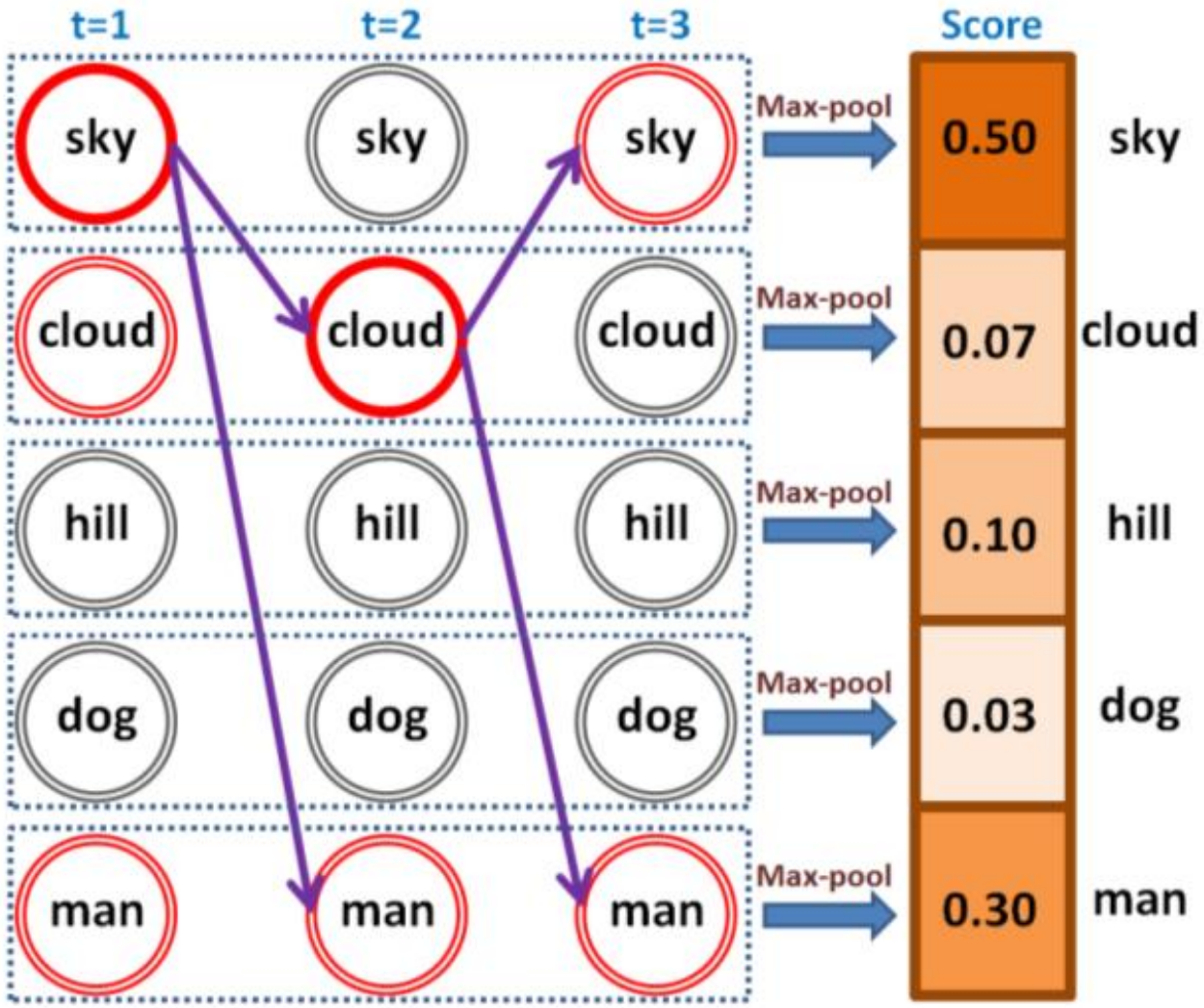


- No specific label order for multi-label image annotation
- Predefined frequency-based label ordering imposed
- Rare to Frequent ordering works the best

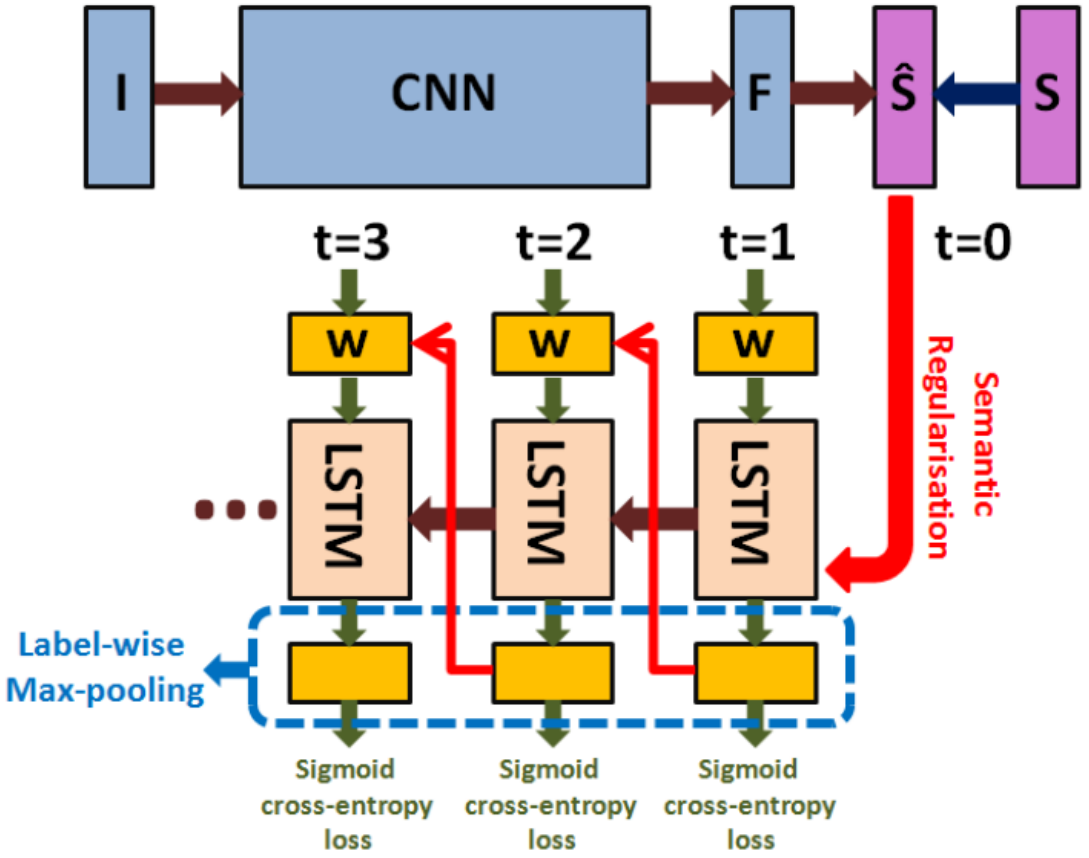
Limitations of existing approaches

- Hard constraint on the model
- Error made in first time step gets propagated to the next step
- Hard to predict the rare label first
- Hard to predict the rare label in the end
- May not reflect the true label dependencies
- Dependent on dataset-specific statistics

Proposed Learning of multiple prediction paths



Proposed Training



y *ground truth binary label vector*

y_t *ground truth binary label vector at time t*

$\hat{y}_t$ *predicted label scores vector*

l_t *hard predicted label*

a_{t-1} *one-hot encoding of hard predicted label in previous time step*

$$l_t = \underset{c \in \{1,2,\ldots,C\}}{\arg\max} \; \hat{y}_t^c \qquad \tilde{a}_t = \neg a_{t-1}$$

$$y_t = \tilde{a}_t \odot y$$

$$\mathcal{L}_t = y_t \cdot log(\sigma(\hat{y}_t)) + (1 - y_t) \cdot log(1 - \sigma(\hat{y}_t))$$

Results

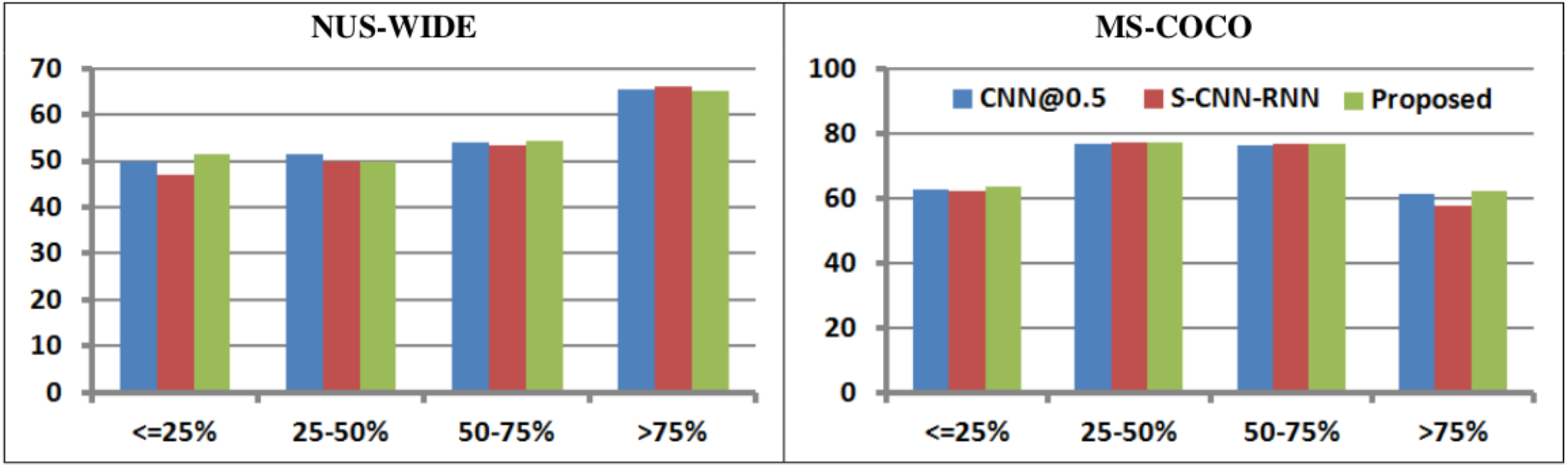
Dataset→	NUS-WIDE						MS-COCO					
Metric→	Per-label			Per-image			Per-label			Per-image		
Method↓	P_L	R_L	$F1_L$	P_I	R_I	$F1_I$	P_L	R_L	$F1_L$	P_I	R_I	$F1_I$
CNN@Top-K	47.66	55.95	48.26	55.66	78.20	59.67	64.59	62.59	61.65	64.07	78.01	64.96
CNN@0.5	68.82	49.01	55.35	80.82	70.58	69.85	81.97	61.97	69.04	87.40	73.96	76.5
CNN-RNN [55]	40.50	30.40	34.70	49.90	61.70	55.20	66.00	55.60	60.40	69.20	66.40	67.80
RIA [25]	52.90	43.60	47.80	68.90	66.70	67.80	64.30	54.10	58.70	74.20	64.50	60.00
Order-Free [7]	59.40	50.70	54.70	69.00	71.40	70.20	71.60	54.80	62.10	74.20	62.20	67.7
S-CNN-RNN [36]	60.35	54.40	54.22	73.22	75.34	71.25	78.89	62.92	68.19	83.03	57.91	64.34
Proposed	60.85	54.43	55.32	76.50	73.06	70.62	77.09	64.32	69.63	84.90	75.83	77.13

Comparison with S-CNN-RNN

NUS-WIDE		MS-COCO	
S-CNN-RNN [36]	Proposed	S-CNN-RNN [36]	Proposed
68.06	84.05	81.44	93.49

Comparison between S-CNN-RNN and the proposed approach based on **top-1 accuracy** of the predicted label (averaged over all the images) on the two datasets

Further Analysis



Comparison between CNN@0.5, S-CNN-RNN and the proposed model for **$F1_L$ scores in different label bins sorted by frequency**

Comparison with CNN

- CNN @TopK
  - Traditional approach
  - Images can have variable no. of labels
- CNN @a
  - Higher precision but reduced recall than CNN@TopK
  - Our model has higher recall, performs better for rare labels
  - First in literature to compare against CNN@a
  - Bound in scope of CNN-RNN performance

Qualitative Results

				
clouds, sky	clouds, sky, water, beach, buildings	clouds, sky, window	ocean, water, waves	animal, dog
clouds, house, sky	clouds, sky, water	clouds, sky, vehicle, window	animal, ocean, water, waves	animal, dog, sky

Conclusion

- Our proposed CNN-RNN model can learn semantic relationships among the labels without the need of providing a pre-defined label ordering.
- CNN-RNN model performs well in terms of recall but is limited by the drop in precision

Future Directions

- CNN-RNN to sequentially look into different image regions
- Learning nouns, verbs, adjectives, "is-a", "has-a" etc. label relationships
- Multi scale region proposal-based methods for small object detection
- Performance impact related to missing labels in test data

Publication

- *Ayushi Dutta, Yashaswi Verma, C. V. Jawahar. Automatic Image Annotation: The Quirks and What Works. In Multimedia Tools and Applications, 2018*

Thank You

Questions ?