Structured Matching for Phrase Localization

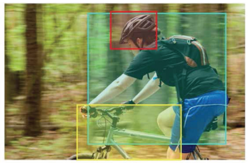
Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, Jia Deng

Computer Science and Engineering, University of Michigan, Ann Arbor

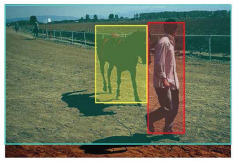
Introduction

Phrase Localization

- Given an image and its textual description, locate the image regions that correspond to the noun phrases in the description.



A woman wearing a black helmet riding on a bike.



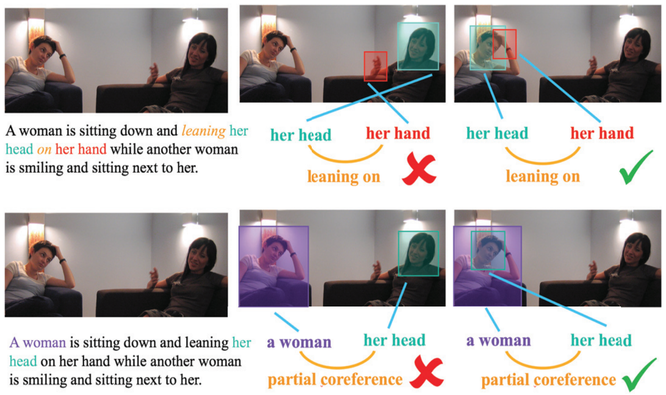
A man is working his horse on a racetrack.

(examples from Flickr30Entities [1] dataset)

Our Contribution

- For the task of phrase localization, we propose a structured matching of phrases and regions that encourages the semantic relations between phrases to agree with the visual relations between regions.
- We formulate structured matching as a discrete optimization problem and relax it to a linear program to enable end-to-end training with neural networks.

Motivation

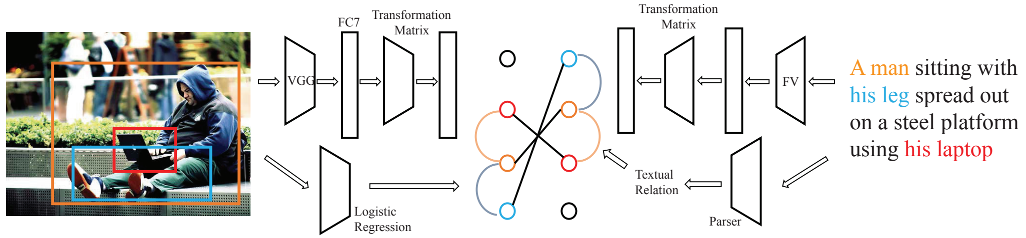


- Phrase localization requires a deep understanding of semantic relations among phrases.
- This leads to the problem of structured matching of regions and phrases:
  (1) individual regions agree with their corresponding phrases.
  (2) visual relations among regions agree with textual relations among corresponding phrases.

Contact

Email: mzwang@umich.edu

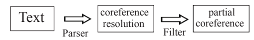
Approach



Partial Coreference

- Partial coreference : introduced by possessive prenouns "his" , "her" or "its". For example:

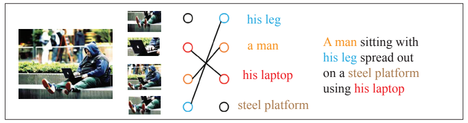
  A woman is dressed in Asian garb with a basket of goods on her hip.
  An instructor is teaching his students how to escape a hold in a self-defense class.



Bipartite Matching

- Denote y_{ij} as a matching configuration, $y_{ij}=1$ is phrase p_i is matched with region r_j, $y_{ij}=0$ otherwise.
- Denote w_{ij} as the weight of phrase p_i and region r_j.
- Solve the bipartite matching as a problem of Linear Programming:

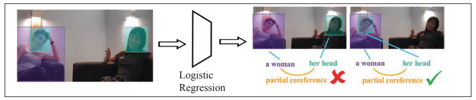
$$\max_y \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}y_{ij}$$
$$\text{s.t.} \sum_{j=1}^{m} y_{ij}=1,\ i=1,2,\dots,n$$
$$\sum_{i=1}^{n} y_{ij}\le 1,\ j=1,2,\dots,m$$
$$0\le y_{ij}\le 1,\ i=1,\dots,n, j=1,\dots,m.$$

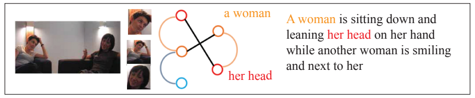


Structured Matching

- Denote z_{ijst} as the joint configuration of phrase p_i, p_s with r_j, r_t.
- Relaxation: Refer y_{ij} as the probability of p_i is matched with r_j. Then z_{ijst} is the joint probability of p_i is matched with r_j, p_s is matched with r_t. With the rule of marginalization:

$$\sum_{t=1}^{m} z_{ijst} = \sum_t \Pr(R(p_i)=r_j, R(p_s)=r_t) = \Pr(R(p_i)=r_j) = y_{ij}$$





$$\max_{y\in\mathcal{Y}} \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}y_{ij} + \lambda \sum_{(i,s)\in Q}^{n}\sum_{j,t}^{m} z_{ijst}\, g(r_j, r_t)$$
$$\text{s.t.} \sum_{j=1}^{m} y_{ij}=1, \text{ for } i=1,2,\dots,n$$
$$\sum_{i=1}^{n} y_{ij}\le 1, \text{ for } j=1,2,\dots,m$$
$$\sum_{t=1}^{m} z_{ijst}=y_{ij} \text{ for any } i,j,s$$
$$\sum_{j=1}^{m} z_{ijst}=y_{st} \text{ for any } i,s,t$$
$$0\le y_{ij}\le 1, \text{ for all } i,j$$
$$0\le z_{ijst}\le 1, \text{ for all } i,j,s,t.$$

Experiments

Experiment Setup

- Dataset: Flickr30K Entities [1].
  - 31783 images and 500k regions.
  - 500k noun phrases and 70k unique phrases.
- Evaluate with Recall@1 across all phrases.
- A region is true if it overlaps with the ground truth in terms of IoU > 0.5.

Results

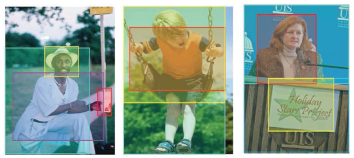
Methods	Accuracy (Recall@1)
CCA [1]	25.30
NonlinearSP [2]	26.70 (43.89)
SCRC [9]	27.80
GroundR [3]	29.02 (47.70)
MCB [28]	(48.69)
CCA [29]	(50.89)
Ours: CCA+Fast-RCNN	39.44
Ours: Matching	41.78
Ours: Structured Matching	42.08

Methods	accuracy (Recall@1) on PC phrases only
Bipartite Matching	47.8
Structured Matching	49.3

Methods	person	clothing	body parts	animals	vehicles	instruments	scene	other
CCA[1]	29.58	24.20	10.52	33.40	34.75	35.80	20.20	20.75
GroundR[3]	44.24	9.93	1.91	45.17	46.00	20.99	30.20	16.12
	(53.80)	(34.04)	(7.27)	(49.23)	(58.75)	(22.84)	(52.07)	(24.13)
CCA[29]	(64.73)	(46.88)	(17.21)	(65.83)	(68.75)	(37.65)	(51.39)	(31.77)
Ours: CCA+FRCN	55.39	32.78	16.25	53.86	48.50	19.14	28.97	23.56
Ours: Bipartite	57.94	34.43	16.44	56.56	51.50	27.16	33.42	26.23
Ours: Structured	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
Upperbound	89.36	66.48	39.39	84.56	91.00	69.75	75.05	67.40

Qualitative Results

Successful cases

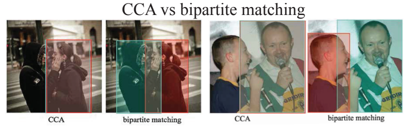


A black man wearing a write suit and hat is holding a paper cup.

A boy wearing an orange shirt is playing on a swing.

A blond woman speaks at a podium labeled Holiday Stars Projectin front of a blue wall.

CCA vs bipartite matching



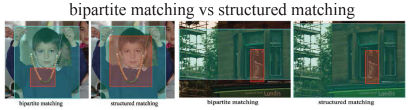
CCA — bipartite matching

A man wearing a black jacket with a woman wearing a black jacket are standing close to each other.

CCA — bipartite matching

A man and a boy holding microphones.

bipartite matching vs structured matching



bipartite matching — structured matching

A young boy shows his brown and green bead necklace.

bipartite matching — structured matching

A man is working on his house by repairing the windows

References

[1] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image to setence models. ICCV 2015

[2] Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embedings.CVPR 2016

[3] Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. ECCV 2016

[9] Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. CVPR 2016

[28] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv 2016

[29] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV 2016

Code

https://github.com/mingzhew/structured-matching