Every Picture Tells a Story: Generating Sentences from Images

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, David Forsyth

University of Illinois at Urbana-Champaign

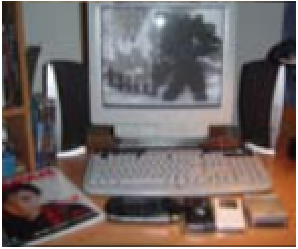
Images most from Farhadi et al. (2010)

Goal

Auto-annotation: find text annotations for images

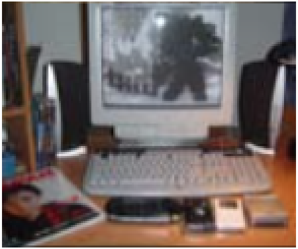
Goal

Auto-annotation: find text annotations for images



Goal

Auto-annotation: find text annotations for images



- This is a lot of technology.
- Somebodys screensaver of a pumpkin
- A black laptop is connected to a black Dell monitor
- This is a dual monitor setup
- Old school Computer monitor with way to many stickers on it

Goal

Auto-illustration: find pictures suggested by given text

Goal

Auto-illustration: find pictures suggested by given text

Yellow train on the tracks.

Goal

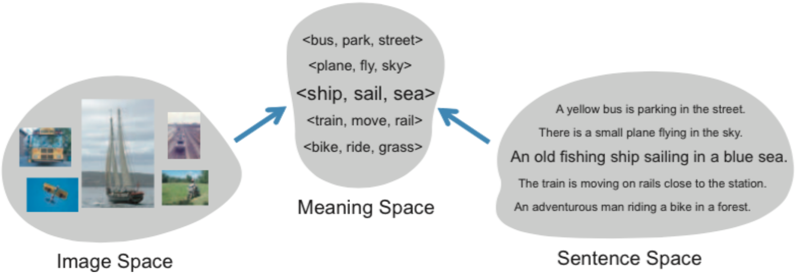
Auto-illustration: find pictures suggested by given text

Yellow train on the tracks.



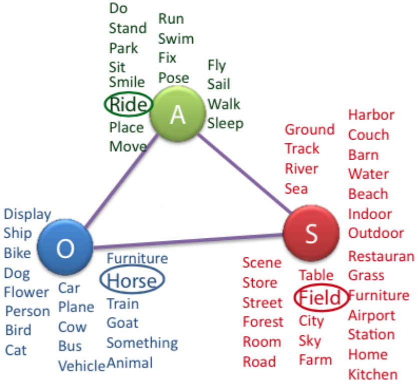
Overview

- Evaluate the similarity between a sentence and an image
- Build around an intermediate representation



Meaning Space

- a triplet of $\langle object, action, scene \rangle$.
- predicting a triplet involves solving a multi-label Markov Random Field

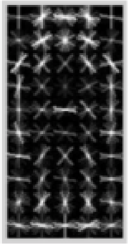


Node Potentials

- Computed as a linear combination of scores from detectors/classifiers

- Image Features
  - DPM response: max detection confidence for each class, their center location, aspect ratio and scale
  - Image classification scores: based on geometry, HOG features and detection response
  - GIST based scene classification: scores for each scene

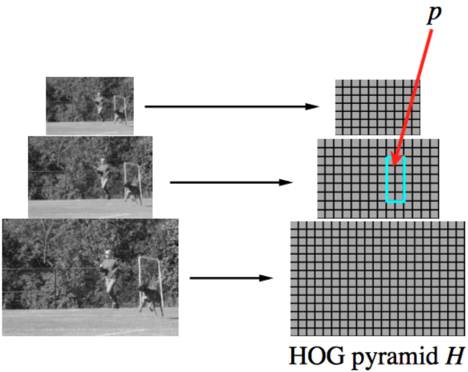
Deformable Part-based Model (DPM)



- Using sliding window approach to search for all possible locations
- Adopt Histogram of Oriented Gradients(HOG) features & linear SVM classifiers

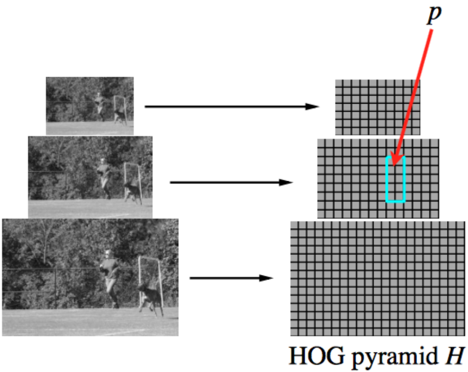
Images from Felzenszwalb et al. (2008)

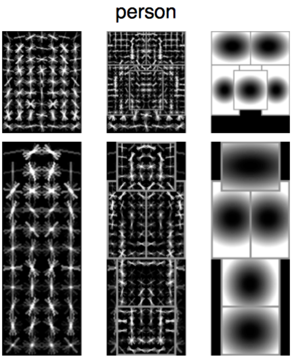
Deformable Part-based Model (DPM)



- Build HOG pyramid thus fix-sized filter can be used

Deformable Part-based Model (DPM)

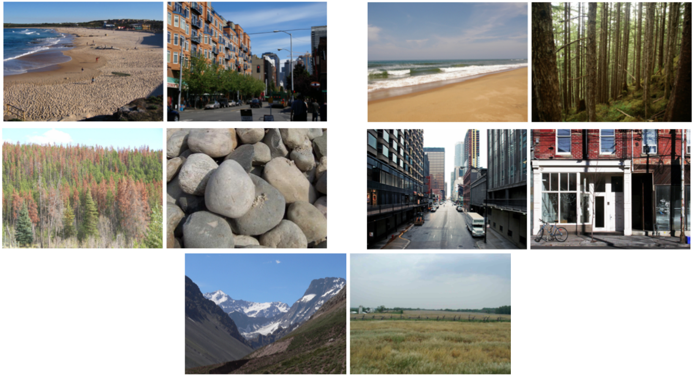




- Build HOG pyramid thus fix-sized filter can be used
- Sum the score from root/part filters and deformation costs

GIST

- Using a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) for scene representation
- Estimate these dimensions from DFT and windowed DFT



Images from Oliva and Torralba (2001)

Node Potentials

- Node features, Similarity Features
- Node features
  - a #-of-nodes-dimensional vector
  - obtained by feeding image features into a linear SVM
- Similarity Features
  - Average of the node features over KNN in the training set to the test image by matching image features
  - Average of the node features over KNN in the training set to the test image by matching those node features

Edge Potentials

- One parameter per edge results in large number of parameters
- Linear combination of multiple initial estimates
- The weights of linear combination can be learnt
  - The normalized frequency of the word A in our corpus, $f(A)$
  - The normalized frequency of the word B in our corpus, $f(B)$
  - The normalized frequency of (A and B) at the same time, $f(A, B)$
  - $\frac{f(A,B)}{f(A)f(B)}$

Sentence Potentials

- Extract (object,action) pairs by Curran & Clark parser.
- Extract head nouns of prepositional phrases etc. for scene
- Use Lin Similarity to determine semantic distance between two words
- Determine actions commonly co-occurring from $8,000$ images captions
- Compute sentence node potentials from these measures
- Estimating edge potentials is identical with that for images

Learning & Inference

- Learn mapping from image space to meaning space
- Learn mapping from sentence space to meaning space

$$\min_{w} \frac{\lambda}{2}||\omega||^2 + \frac{1}{n} \sum_{i \in examples} \xi_i$$

$$s.t. \quad \forall i \in examples:$$

$$\omega\Phi(x_i, y_i) + \xi_i \geq \max_{y \in meaningspace} \omega\Phi(x_i, y) + L(y_i, y)$$

$$\xi_i \geq 0$$

Learning & Inference

- Search for the best triplet that maximizes

$$\arg\max_{y} \omega^T \Phi(x_i, y)$$

- A multiplicative model prefer all response to be good

$$\arg\max_{y} \prod \omega^T \Phi(x_i, y)$$

- Greedily relax an edge, solving best path and re-scoring

Matching

- Match sentence triplets and image triplets
- Obtain top k ranking triplets from sentence, compute their ranks as image triplet
- Obtain top k ranking triplets from image, compute their ranks as sentence triplet
- Sum the ranks of all these sets

Text Information and Similarity measure is used to take care of out of vocabulary words that occurs in sentences but are not being learnt by a detector/classifier

Evaluation

- Build dataset with images and sentences from PASCAL 2008 images
- Randomly select 50 images per class (20 class in total)
- Label 5 sentences per image on AMT
- Manually add labels for triplets of $\langle \textit{objects}, \textit{actions}, \textit{scenes} \rangle$
- Select 600 images for training and 400 for testing

Measures:

- Tree-F1 measure:
  - Build taxonomy tree for objects, actions and scenes
  - Calculate F1 score for precision and recall
  - Tree-F1 score is the mean of F1 scores for objects, actions and scenes
- BLUE score:
  - Measure if the generated triplet appear in the corpus or not

Results

Mapping images to meaning space

	Obj	No Edge	FW(A)	SL(A)	FW(M)	SL(M)
Mean Tree-F1 for first 5	0.44	**0.52**	0.38	0.45	0.47	0.51
Mean BLUE for first 5	0.24	0.27	0.16	0.58	**0.76**	0.74
Mean Tree-F1 for first 5 objects	**0.59**	0.58	0.36	0.53	0.55	0.57
Mean Tree-F1 for first 5 actions	0.27	**0.52**	0.50	0.37	0.42	0.47
Mean Tree-F1 for first 5 scenes	0.28	0.48	0.28	0.44	0.46	**0.48**

Table 1. Evaluation of mapping from the image space to the meaning space. "Obj" means when we only consider the potentials on the object node and use uniform potentials for other nodes and edges. "No Edge" means assuming a uniform potential over edges. "FW(A)" stands for fixed weights with additive inference model. This is the case where we use all the potentials but we don't learn any weights for them. "SL(A)" means using structure learning with additive inference model. "FW(M)" is similar to "FW(A)" with the exception that the inference model is multiplicative instead of additive. "SL(M)" is the structure learning with multiplicative inference.

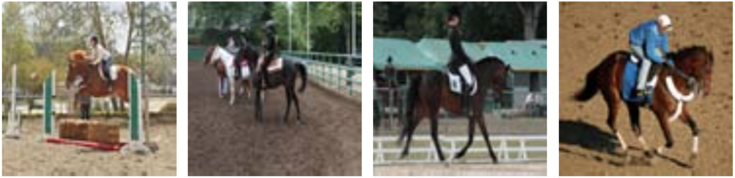
Results: Auto-annotation

	(pet, sleep, ground) (dog, sleep, ground) (animal, sleep, ground) (animal, stand, ground) (goat, stand, ground)	see something unexpected. Cow in the grassfield. Beautiful scenery surrounds a fluffly sheep. Dog hearding sheep in open terrain. Cattle feeding at a trough.
	(furniture, place, furniture) (furniture, place, room) (furniture, place, home) (bottle, place, table) (display, place, table)	Refrigerator almost empty. Foods and utensils. Eatables in the refrigerator. The inside of a refrigerator apples, cottage cheese, tupperwares and lunch bags. Squash apenny white store with a hand statue, picnic tables in front of the building.
	(transportation, move, track) (bike, ride, track) (transportation, move, road) (pet, sleep, ground) (bike, ride, road)	A man stands next to a train on a cloudy day A backpacker stands beside a green train This is a picture of a man standing next to a green train There are two men standing on a rocky beach, smiling at the camera. This is a person laying down in the grass next to their bike in front of a strange white building.
	(display, place, table) (furniture, place, furniture) (furniture, place, furniture) (bottle, place, table) (furniture, place, home)	This is a lot of technology. Somebody's screensaver of a pumpkin A black laptop is connected to a black Dell monitor This is a dual monitor setup Old school Computer monitor with way to many stickers on it

Results: Auto-illustration

A two girls in the store.

A horse being ridden within a fenced area.



Failure Case

	A male and female giving pose for camera. A peaceful garden The food is ready on table.
	The two girls read to drive big bullet. Man with a goatee beard kneeling in front of a garden fence. Lone bicyclist sitting on a bench at a snowy beach.
	Black goat in a cage Horse behind a fence Wooly sheep standing next to a fence on a sunny day.

Discussion

- Sentences are not generated, but searched from a pool of candidate sentences

Discussion

- Sentences are not generated, but searched from a pool of candidate sentences
- Using triplet limits the representation of meaning space

Discussion

- Sentences are not generated, but searched from a pool of candidate sentences
- Using triplet limits the representation of meaning space
- Proposed dataset is small

Discussion

- Sentences are not generated, but searched from a pool of candidate sentences
- Using triplet limits the representation of meaning space
- Proposed dataset is small
- Using Recall@K and median rank as performance measure

Summary

- Proposes a system to compute score linking of an image to a sentence and vice versa
- Evaluates their methodology on a novel dataset consisting of human-annotated images (PASCAL Sentence Dataset)
- Quantitative evaluation on the quality of the predictions

A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.