Visual Recognition with Text (CSC2523)
Projects

Proposal due: Feb 15, 2015, 11.59pm
Final report due: March 28, 2015, 11.59pm
Presentation: last day of class (April 1)

As part of project work, each student needs to:

- Write a project proposal
- Write a project report
- Present his/her project work in class

By Feb 15 you will need to hand in the **project proposal**. The proposal should have a description of the problem you'll be working on and at least initial ideas of how you want to tackle it. The **final report**, which should include problem definition and motivation, literature overview, and a detailed description of your work including results, is due on March 28. You can submit the project proposal and final report through CDF. The **presentations** will be scheduled on the last day of class.

Projects can be done individually or in pairs. If done in a pair, the report should reflect what each student contributed to the project.

The grade will take into account the following factors:

- The difficulty of the problem
- Your ideas to tackle a problem: how appropriate the techniques you chose are for the problem. Coming up with novel ideas is obviously a big plus.
- Your implementation: how far down the project you arrived
- Novelty
- Thoroughness of your report: How well you describe and motivate the problem, review related work, clarity and quality of the approach, how well you analyzed your results.

If you are tackling a difficult problem, it may happen that you will not finish the project in due date. That's absolutely fine as long as what you have done is reasonable.

You can choose to implement a paper of your choice (no novelty), extend an existing paper, propose and work on a new topic/problem, tackle an existing problem in a new way. The grade A+ will only be given to students who's novelty and quality of the project is exceptional. If you decide to work on description generation in the image domain, you are highly encouraged to compete on the Microsoft's CoCo challenge (given that it will be released in time).

Topics

Here is a list of possible topics. Please see the list of papers below as (non-exhaustive) references, as well as the references within papers for a more complete list. You can also come up with your own topic as long as it's related to the images and text domain. If you're not sure or if you want to discuss some topic/problem send Sanja an email (fidler@cs.toronto.edu).

- Description generation from an image/video
- Image generation from description
- Learning visual concepts (objects, actions, attributes, etc) from images/videos annotated (only) with lingual descriptions
- Aligning descriptions with images or videos
- Learning concepts from questions and answers
- Word-sense disambiguation
- Infer high-level semantics in images/videos via e.g. web-mined text or online knowledge databases (e.g. conceptnet or wikipedia)
- Improving visual parsing by exploiting image/video descriptions
- Visual search based on complex lingual queries
- Detecting which words in sentences (paired with images) are visual
- Text detection in the wild or application driven
- Joint parsing of images and text (e.g. solving coreference resolution based on image information and text)
- Efficient crowd-sourcing of large image and text datasets (e.g. via games)
- Indoor/outdoor navigation via lingual descriptions

Paper List

Below is a long (but not exhaustive) list of papers on these topics:

1. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
   R. Kiros, R. Salakhutdinov, R. S. Zemel
   (arXiv:1411.2539), Nov 2014
   Topic(s): Description generation (image), neural netwoks

2. Deep Visual-Semantic Alignments for Generating Image Descriptions
   A. Karpathy, L. Fei-Fei
   (arXiv:1412.2306), Dec 2014
   Project page: http://cs.stanford.edu/people/karpathy/deepimagesent/
   Topic(s): Description generation (image), image-to-text alignment, neural netwoks

3. Learning a Recurrent Visual Representation for Image Caption Generation
Xinlei Chen, C. Lawrence Zitnick
(arXiv:1411.5654), Nov, 2014
Topic(s): Description generation (image), neural networks

4. From Captions to Visual Concepts and Back
Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollr, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, Geoffrey Zweig
(arXiv:1411.4952), Nov 2014
Topic(s): Description generation (image), neural networks

5. Long-term Recurrent Convolutional Networks for Visual Recognition and Description
Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell
(arXiv:1411.4389), Nov 2014
Topic(s): Description generation (image), neural networks

6. Show and Tell: A Neural Image Caption Generator
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan
(arXiv:1411.4555), Nov 2014
Topic(s): Description generation (image), neural networks

7. Explain Images with Multimodal Recurrent Neural Networks
Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Alan L. Yuille
(arXiv:1410.1090), Oct 2014
Topic(s): Description generation (image), neural networks

8. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics
M. Hodosh, P. Young, J. Hockenmaier
Journal of Artificial Intelligence Research, 2013
Project page: `http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/KCCA.html`
Topic(s): Description generation (image), discussion on evaluation metrics

9. Every Picture Tells a Story: Generating Sentences for Images
A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. A. Forsyth
ECCV, 2010
Topic(s): Description generation (image)

10. Im2Text: Describing Images Using 1 Million Captioned Photographs
V. Ordonez, G. Kulkarni, T. L. Berg
NIPS, 2011
Project page: `http://vision.cs.stonybrook.edu/~vicente/sbucaptions/`
Topic(s): Description generation (image)

11. TREETALK: Composition and Compression of Trees for Image Descriptions
P. Kuznetsova, V. Ordonez, T. L. Berg, Y. Choi
TACL, 2014
Topic(s): Description generation (image)

12. How many words is a picture worth? Automatic caption generation for news images
    Y. Feng, M. Lapata
    ACL, 2010
    Topic(s): Description generation (image), news images

13. Translating Video Content to Natural Language Descriptions
    M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele
    ICCV, 2013
    Topic(s): Description generation (video)

14. Video In Sentences Out
    A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, Z. Zhang
    UAI, 2012
    Project page: `https://engineering.purdue.edu/~qobi/mindseye/`
    Topic(s): Description generation (video)

15. Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos
    A. Gupta, P. Srinivasan, J. Shi, L. S. Davis
    CVPR 2009
    Topic(s): Description generation (video), storyline model

16. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition
    S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko
    ICCV, 2013
    Topic(s): Description generation (video)

17. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching
    P. Das, C. Xu, R. F. Doell, and J. J. Corso
    CVPR, 2013
    Topic(s): Description generation (video)

18. Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora
    R. Socher, L. Fei-Fei
    CVPR, 2010
    Topic(s): Learning visual models from text (image domain)

19. Beyond Nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers
    A. Gupta, L. S. Davis
    ECCV, 2008
    Topic(s): Learning visual models from text (image domain)

20. Matching Words and Pictures
    K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, M. I. Jordan

JMLR, 2003
Topic(s): Learning visual models (segmentation) from text (image domain)

21. Discovering Hierarchical Object Models from Captioned Images
M. Jamieson, Y. Eskin, A. Fzaly, S. Stevenson, S. Dickinson
CVIU, 2012
Topic(s): Learning visual models from text (image domain)

22. Learning Everything about Anything: Webly-Supervised Visual Concept Learning
S. K. Divvala, A. Farhadi, C. Guestrin
CVPR, 2014
Project page: `http://levan.cs.washington.edu/`
Topic(s): Learning visual concepts from the web

23. Video Event Understanding using Natural Language Descriptions
V. Ramananthan, P. Liang, L. Fei-Fei
ICCV, 2013
Topic(s): Learning visual models from text (video domain)

24. Grounded Language Learning from Video Described with Sentences
H. Yu, J.M. Siskind
ACL, 2013
Topic(s): Learning visual models from text (video domain)

25. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework
L.-J. Li, R. Socher, L. Fei-Fei
CVPR, 2009
Topic(s): Learn visual models from (noisy) tags (image domain)

26. DeViSE: A Deep Visual-Semantic Embedding Model
A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M.'A. Ranzato, T. Mikolov
NIPS, 2013
Topic(s): Learn visual models via text

27. Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions
M. Elhoseiny, B. Saleh, A. Elgammal
ICCV, 2013
Topic(s): Zero-shot learning via text

28. Improving Video Activity Recognition using Object Recognition and Text Mining
T. S. Motwani, R. J. Mooney
ECAI, 2012
Topic(s): Learning visual models from text (video domain), video parsing via text

29. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input
M. Malinowski, M. Fritz
NIPS, 2014
Project page: `https://www.d2.mpi-inf.mpg.de/visual-turing-challenge`
Topic(s): Question and answering

30. Inferring the Why in Images
H. Pirsiavash, C. Vondrick, A. Torralba
TR, 2014
Project page: http://web.mit.edu/why/
Topic(s): Learning high-level visual semantics via text

31. Detecting Visual Text
J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. Daum III, A. C. Berg, T. L. Berg
NAACL 2012
Topic(s): Detecting what's visual in descriptions

32. Learning the Visual Interpretation of Sentences
C. L. Zitnick, D. Parikh, L. Vanderwende
ICCV, 2013
Topic(s): Image generation from descriptions

33. WordsEye: An Automatic Text-to-Scene Conversion System
Bob Coyne, Richard Sproat
SIGGRAPH, 2001
Topic(s): Image generation from descriptions

34. Joint person naming in videos and coreference resolution in text
V. Ramanathan, A. Joulin, P. Liang, L. Fei-Fei
ECCV, 2014
Topic(s): Joint image and text modeling

35. Movie/Script: Alignment and Parsing of Video and Text Transcription
T. Cour, C. Jordan, E. Miltsakaki, B. Taskar
ECCV, 2008
Topic(s): Text-to-video alignment

36. Aligning Plot Synopses to Videos for Story-based Retrieval
M. Tapaswi, M. Baeuml, R. Stiefelhagen
International Journal of Multimedia Information Retrieval (IJMIR), 2014
Project page: https://cvhci.anthropomatik.kit.edu/~mtapaswi/projects/story_based_retrieval.html
Topic(s): Text-to-video alignment

37. Visual Semantic Search: Retrieving Videos via Complex Textual Queries
D. Lin, S. Fidler, C. Kong, R. Urtasun
CVPR 2014
Topic(s): Visual search with complex text queries

38. A Sentence is Worth a Thousand Pixels
S. Fidler, A. Sharma, R. Urtasun
CVPR, 2013
Topic(s): Improving image parsing via text

39. What are you talking about? Text-to-Image Coreference
C. Kong, D. Lin, M. Bansal, R. Urtasun, S. Fidler
CVPR, 2014
Topic(s): Improving image (RGB-D domain) parsing via text, joint image and text models (for coreference resolution)

40. Robots with Language: Multi-Label Visual Recognition Using NLP
Y. Yang, C. L. Teo, C. Fermuller, Y. Aloimonos
ICRA, 2013
Topic(s): Improving visual parsing (video) via text, robotics

41. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning
B. Siddiquie, A. Gupta
CVPR 2010
Topic(s): Active learning using lingual questions

42. ReferItGame: Referring to Objects in Photographs of Natural Scenes
S. Kazemzadeh, V. Ordonez, M. Matten, T. L. Berg
EMNLP, 2014
Project page: http://tamaraberg.com/referitgame/
Topic(s): Crowd-sourcing image and text data via fun games

43. Adopting Abstract Images for Semantic Scene Understanding
C. L. Zitnick, Member, R. Vedantam, D. Parikh
TPAMI, 2015
Topic(s): Crowd-sourcing fun image and text data (cartoon images)

44. Understanding and Predicting Importance in Images
A. C. Berg, T. L. Berg, H. Daum III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, K. Yamaguchi
CVPR 2012
Project page: http://tamaraberg.com/importanceDataset/
Topic(s): Analysis of what people describe in images

45. Unsupervised Learning of Visual Sense Models for Polysemous Words
K, Saenko, T. Darrell
NIPS, 2008
Topic(s): Word/image-sense disambiguation

46. Joint Image and Word Sense Discrimination for Image Retrieval
A. Lucchi, J. Weston
ECCV, 2012
Topic(s): Word/image-sense disambiguation

47. Deep Features for Text Spotting
M. Jaderberg, A. Vedaldi, A. Zisserman
ECCV, 2014
Project page: http://www.robots.ox.ac.uk/~vgg/publications/2014/Jaderberg14/
Topic(s): Detecting text in images

48. End-to-End Scene Text Recognition
K. Wang, B. Babenko, S. Belongie
ICCV, 2011
Project page: http://vision.ucsd.edu/~kai/svt/
Topic(s): Detecting text in images (Google's Street View)

49. Tell Me Dave: Context-Sensitive Grounding of Natural Language to Mobile Manipulation Instructions
D. K Misra, J. Sung, K. Lee, A. Saxena
RSS, 2014
Project page: `http://tellmedave.com/`
Topic(s): Learning lingual commands, robotics

50. Framework for Learning Semantic Maps from Grounded Natural Language Descriptions
Walter, M.R., Hemachandra, S., Homberg, B., Tellex, S., Teller, S., A
International Journal of Robotics Research, 2014
Topic(s): Learning environment maps from lingual descriptions, robotics

51. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation
S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, N. Roy
AAAI 2011
Topic(s): Learning lingual commands, robotics

52. Parsing Natural Scenes and Natural Language with Recursive Neural Networks
R. Socher, C. Lin, A. Y. Ng, C. Manning
ICML, 2011
Topic(s): Parsing of images and sentences