Project Proposal

Exploring Domain Adaptation for the Image Description Task

Glebys Gonzalez, Maria L. Pacheco
School of Industrial Engineering, Department of Computer Science
gonza337@purdue.edu, pachecog@purdue.edu

October 2, 2015

1 Introduction

In recent years there has been a great effort and ongoing work on systems that associate images with natural language descriptions of them. A number of tasks associated with this problem have been proposed, including caption generation Xu et al. [2015] Chen and Lawrence Zitnick [2015] Karpathy and Li [2014], image search and description as a ranking task Hodosh et al. [2013], cross-modal retrieval Rasiwasia et al. [2010] Costa Pereira et al. [2014] and even image generation Zitnick et al. [2013]. Recently, researchers have asked the question: *are state-of-the-art methods really capturing the meaning of these images?*

Despite all advances and improvement measures that have been achieved in different tasks, there is a general interest in capturing the real semantic meaning of images through their descriptions. Following this intuition, better and richer datasets are being proposed to progress in this direction. The Flickr30k dataset has been expanded with bidirectional annotations in images and sentence descriptions that correlate mentions of the same entities, both in written form and as a specific segment of a picture Plummer et al. [2015]. Microsoft released a couple of datasets for this problem also, the COCO Dataset, which also provides richer annotation to their images Lin et al. [2014], and the Abstract Image Dataset, a collection of clipart images with the intention of abstracting from the characteristics of real images and providing a simpler ground to focus on semantic meaning understanding Zitnick and Parikh [2013].

The idea of using abstract images is interesting and, assuming that their intuition is correct, it is even more interesting if we could take the extracted knowledge for this idealized setting and transfer it to real scenarios. For this project we propose to evaluate domain adaptation for this setting and explore the potential of using a simplified abstract domain to improve semantic meaning understanding in real images.

2 Related work

Describe any previous work you found related to your project

3 Problem formulation

The problem of text based image description has been formulated as a ranking task Hodosh et al. [2013]. Taking the set of images and their descriptions, we can evaluate for each description how well the proposed system ranks an image in contrast to all other images in the set. This framework allows for the analogous task to be evaluated as well, testing how well the system ranks the captions of a given image against the complete caption set.

For both tasks in this cross-modal evaluation the input objects are pairs of the form (i, C) where i corresponds to an image and C corresponds to a set of sentences describing such image. Depending on the task, the pairing would be done from image to captions or from captions to images.

Metrics can be computed automatically, measuring the recall at different levels. For example, the rate of queries in which the correct response was among the top k results and the median rank of the correct responses Hodosh et al. [2013].

Different approaches for this task have tried different feature representations. We intend to evaluate them further to propose an appropriate feature representation for this project.

Microsoft Research released a dataset of abstract scenes and descriptions following the intuition that extracting high-level semantic meaning from real images tends to be difficult. Real images possess a great amount of detail and complex features that are not necessarily needed to capture its semantic meaning.

We propose to explore domain adaptation in this scenario, having two sets: the abstract dataset as the source set and a repository of real images as the target set. The main goal is to learn the correlation of images and sentences in this simplified domain and map it to the more expressive domain of real images.

4 Data and Evaluation plan

As a domain adaptation problem, we need to define our source and target datasets.

- **Source dataset:** The Abstract Scenes Dataset released by Microsoft Research (Clipart). The first version of this dataset contains 10,000 abstract images of kids playing in a park. The images are composed by a set of 80 pieces of clipart representing 58 different objects. These images are arranged in groups of 10, displaying different pictorial versions of the same scene text description. A newer version of the dataset has been released expanding to 6 the number of sentences associated with each image, amounting for a total of 60,000 sentences Zitnick and Parikh [2013].

- **Target dataset (op1):** The Flickr30k Entities dataset is an extension of the Flick30k released by University of Illinois at Urbana Champaign. This dataset contains 30,000 images and 158,000 captions. The dataset has been expanded with an additional 244,000 coreference chains linking mentioned entities in the captions the segment of the images that frame those entities Plummer et al. [2015].

- **Target dataset (op2)**: COCO is a dataset released by Microsoft Research containing more than 300,000 images and 5 captions per image. This dataset also includes object segmentation in images from a total of 80 objects categories Lin et al. [2014].

To evaluate our proposal we intend to take as a baseline the training and retrieval done by only using repositories of real images. Since we intend to measure the advantages of including simpler images in the source domain, the evaluation would focus on the improvements achieved by using the abstract images dataset as a source domain and exploring domain adaptation on the task of image retrieval on real images. All tests are to be done using the same algorithm.

References

Xinlei Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. June 2015.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *Transactions of Pattern Analysis and Machine Intelligence*, 36(3):521–535, March 2014.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, May 2013. ISSN 1076-9757.

Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015.

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval. In *ACM International Conference on Multimedia*, pages 251–260, 2010.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.

C. L. Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

C.L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1681–1688, Dec 2013. doi: 10.1109/ICCV.2013.211.