End to End Scene Graph Generation Using Deep Neural Networks - Towards Image Understanding

Satoshi Tsutsui stsutsui@indiana.edu
Manish Kumar kumar20@umail.iu.edu

1 Project description

Image understanding by computer is advancing exponentially these days due to the phenomenal success of deep learning, but there is still much work left for the computers to reach human level perception. Image classification (sometimes with localization) is one of the standard task, but this is far from the image understanding. The other tasks such as image caption generation or visual question answering have also reached to practical level of quality, but these are still far from the complete image understanding. Image caption generation, which is a task that generates a summary sentence from an image, cannot fully describe rich scenery in an image. Visual question answering also answers simple questions, but cannot answer questions that requires complex reasoning. In order to fully understand an image, we need to know; what objects are in the image, what are the characteristics of objects, and what are the relations between these objects.

With this motivation, Stanford computer vision lab released a dataset called "Visual Genome" \cite { krishnavisualgenome } orhttps://visualgenome.org. This is a dataset that has 100K annotated images with regional descriptors, objects, attributes, relations, and question answering. This is currently the largest and richest dataset available for image understanding.

In this project, we will particularly focus on objects, attributes, and relations, which will be the scene graph eventually. **Our goal is to build a system that takes an arbitrary daily life image and generates its scene graph as an output.**

*scene graph example is attached on the last page as PDF, which is taken from \cite{krishnavisualgenome}.

2 Reading List

2.1 Papers on caption or regional descriptor generation based on neural nets:

```
@article{kiros2014unifying,
 title={Unifying visual-semantic embeddings with multimodal neural language models},
 author={Kiros, Ryan and Salakhutdinov, Ruslan and Zemel, Richard S},
 journal={arXiv preprint arXiv:1411.2539},
 year={2014}
}

@inproceedings{vinyals2015show,
 title={Show and tell: A neural image caption generator},
 author={Vinyals, Oriol and Toshev, Alexander and Bengio, Samy and Erhan, Dumitru},
 booktitle={Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition},
 pages={3156--3164},
 year={2015}
}

@article{xu2015show,
 title={Show, attend and tell: Neural image caption generation with visual attention},
 author={Xu, Kelvin and Ba, Jimmy and Kiros, Ryan and Courville, Aaron and Salakhutdinov, Ruslan and Zemel,
Richard and Bengio, Yoshua},
 journal={arXiv preprint arXiv:1502.03044},
 year={2015}
}
```

@InProceedings{Karpathy_2015_CVPR,
author = {Karpathy, Andrej and Fei-Fei, Li},
title = {Deep Visual-Semantic Alignments for Generating Image Descriptions},
booktitle = {The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2015}
}

2.2 Papers on scene graph

@inproceedings{schuster2015generating,
title={Generating semantically precise scene graphs from textual descriptions for improved image retrieval},
author={Schuster, Sebastian and Krishna, Ranjay and Chang, Angel and Fei-Fei, Li and Manning, Christopher D},
booktitle={Proceedings of the Fourth Workshop on Vision and Language},
pages={70--80},
year={2015}
}
@inproceedings{johnson2015image,
title={Image retrieval using scene graphs},
author={Johnson, Justin and Krishna, Ranjay and Stark, Michael and Li, Li-Jia and Shamma, David A and Bernstein, Michael S and Fei-Fei, Li},
booktitle={Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on},
pages={3668--3678},
year={2015},
organization={IEEE}
}
@inproceedings{krishnavisualgenome,
title={Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations},
author={Krishna, Ranjay and Zhu, Yuke and Groth, Oliver and Johnson, Justin and Hata, Kenji and Kravitz, Joshua and Chen, Stephanie and Kalanditis, Yannis and Li, Li-Jia and Shamma, David A and Bernstein, Michael and Fei-Fei, Li},
year = {2016},
}
@incollection{farhadi2010every,
title={Every picture tells a story: Generating sentences from images},
author={Farhadi, Ali and Hejrati, Mohsen and Sadeghi, Mohammad Amin and Young, Peter and Rashtchian, Cyrus and Hockenmaier, Julia and Forsyth, David},
booktitle={Computer Vision--ECCV 2010},
pages={15--29},
year={2010},
publisher={Springer}
}

3 Research plan and time-line

3.1 Baseline

We will use recurrent neural net based image caption generator that is publicly available [1,2]. Then apply scene graph generator for the generated sentences. The graph generator from sentence is also on public [3].

In addition, we might want to train caption generator from scratch using Visual Genome dataset because available caption generator is trained on MSCOCO, not Visual Genome.

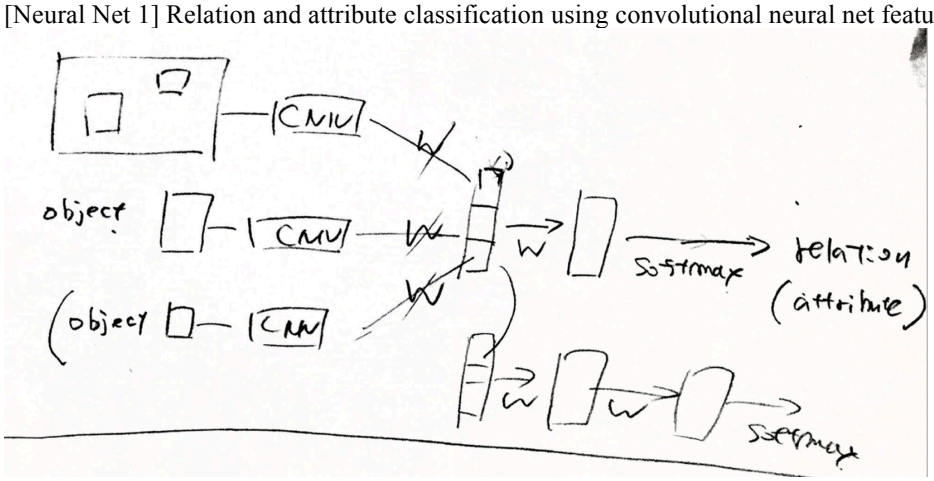
[1]Python implementation: https://github.com/apple2373/chainer_caption_generation
[2]Torch Implementation: https://github.com/karpathy/neuraltalk2

[3]Java implementation: http://nlp.stanford.edu/software/scenegraph-parser.shtml

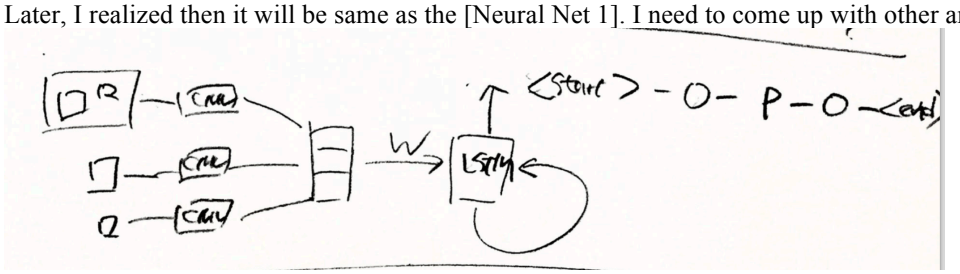
3.2 End to End neural net

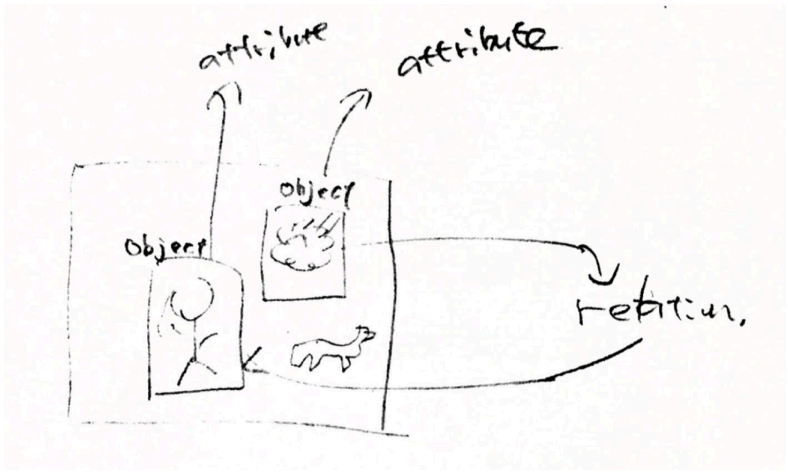
We will also train neural network that generates scene graph directly.
We have two ideas for the neural network architecture.

[Neural Net 1] Relation and attribute classification using convolutional neural net feature



[Neural Net 2] Triplet component generation using recurrent network fed by convolutional neural net feature
Crandall's comment: RNN is overwork. If there is always three elements, we can just use three different output layer.
Later, I realized then it will be same as the [Neural Net 1]. I need to come up with other architecture.





3.3 Time Line

Done by the end of Spring break (Mar 21)

To get familiar with data. (both)
Preprocess data. (satoshi)
Build a base line using [1,2,3]. (satoshi)
To get familiar with (fast or faster) R-CNN to detect object with localization. (manish)
Write evaluation program. Evaluate based on baseline. (manish)

Done by March 31:

Implement [Neural Net 1] with a deep learning framework. Start training on GPU. (both)
Train caption generation model with regional localization. (optional) (satoshi)

Done by April 11:

Evaluate performance based on [Neural Net 1]. (manish)
Modify network (satoshi)

Done by April 18

One more experiment with some network if possible. (satoshi)
Start finishing up the whole system implementation with organized abovementioned modules. i.e. implement a method or class that takes image and output the graph. (manish)

Done by April 29.

Finish up the whole system implementation with organized abovementioned modules. i.e. implement a method or class that takes image and output the graph. (both)
Finish writing paper. (both)

4 Plan for data and experiments

Data: We will use visual genome data. Scene graph data is not yet available in public so we contacted the project member and got almost half of the graph. So we have at least 50K complete dataset. We will split 40K, 0.5K, and 0.5K for training, validation, and test.

Evaluation: Evaluation could be difficult, but current plan is break the graph into the basic triplet components (object - -relation - object) and evaluate precision and recall.

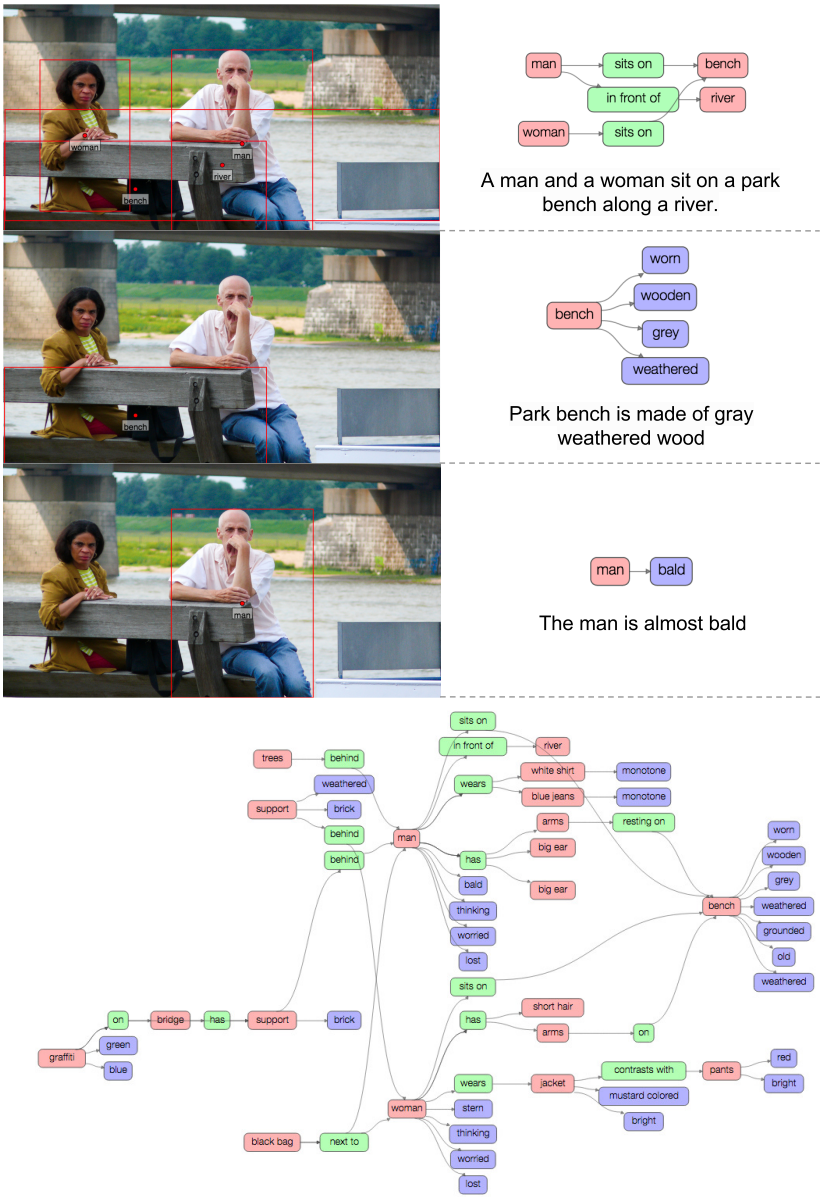


Fig. 2: An example image from the Visual Genome dataset. We show 3 region descriptions and their corresponding region graphs. We also show the connected scene graph collected by combining all of the image's region graphs. The top region description is "a man and a woman sit on a park bench along a river." It contains the objects: `man`, `woman`, `bench` and `river`. The relationships that connect these objects are: *sits_on*(*man*, *bench*), *in_front_of*(*man*, *river*), and *sits_on*(*woman*, *bench*).