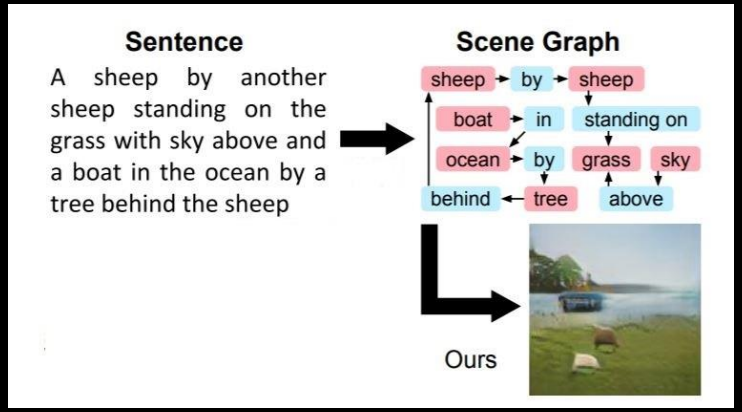
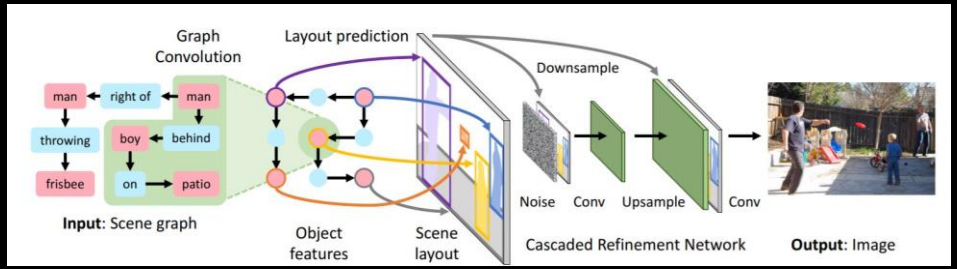
Image Generation from Scene Graphs

Presenter: Chris Rockwell
Authors:
Justin Johnson, Agrim Gupta, Fei-Fei Li

Image Generation from Scene Graphs

- From flexible scene graph, generate representative image
  - Interesting, ambitious: attempts to move towards model building rather than pattern recognition
- Method combines NLP embedding method, graph convolution, CNN, MLP and discriminator to sequentially produce image



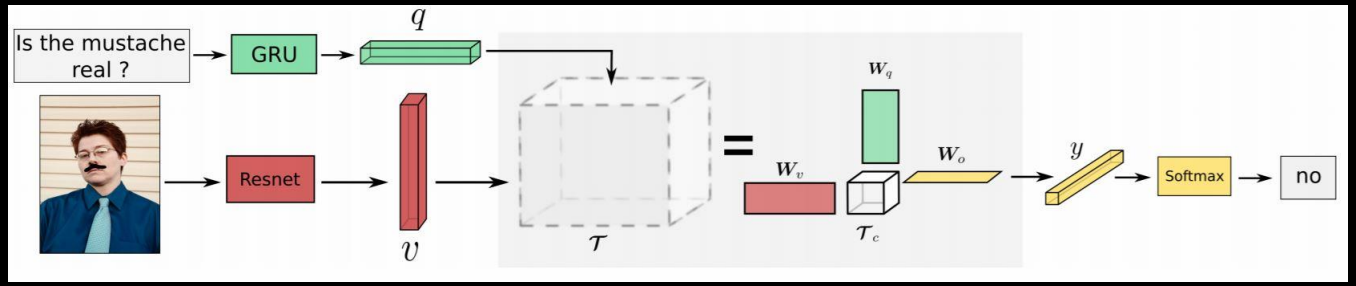


[1] Image credit: Johnson et al., *Image generation from scene graphs*. In CVPR 2018.

Replication & Visual Question Answering

Train model adopting scene graph Github code

1. Produce qualitative results to compare to theirs presented in the paper
2. Answer questions based on images [1] : ground truth, from our trained model, and from their model
    - Due to their quantitative methods requiring human evaluation, or being very limited (bounding box, inception score)



[1] Ben-Younes et al., *Mutan: Multimodal tucker fusion for visual question answering*. In CVPR 2017.

Evaluation Pipeline

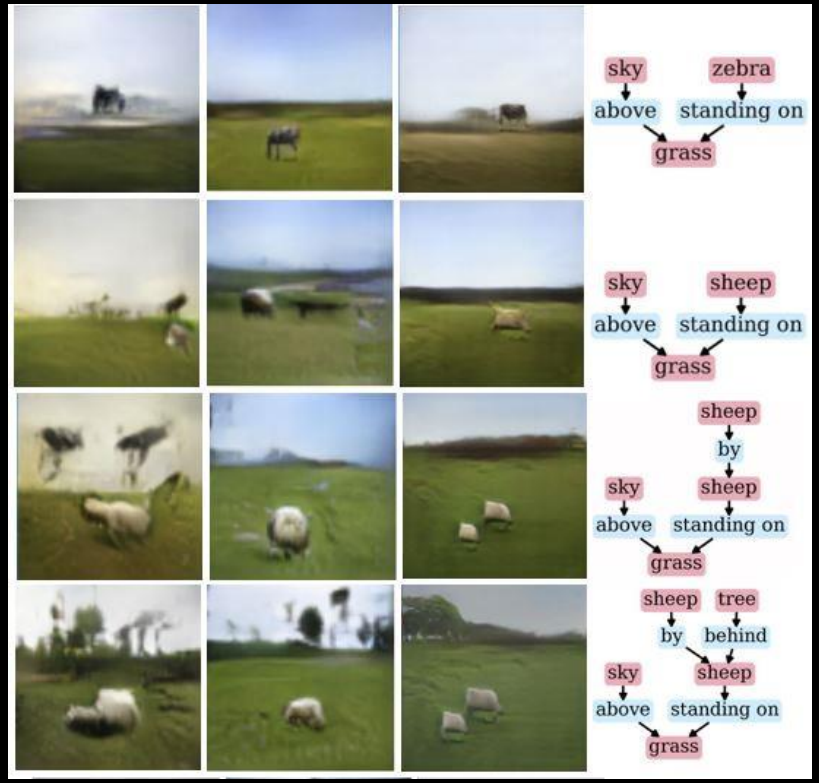
1. Load data
   a) Select 500 compatible images from Visual Genome test set
   b) Get question vector and answer integer from visual question answering model
2. Generate images using authors' and my model, extract features
3. Answer questions using pretrained VQA model
4. Compute accuracy as % of correct answers (2000 classes)

Majority of project

- VQA setup is extensive: extract features, question / answers, datasets
- Compatibility across models: pre and post-processing, hashing of IDs, taking pieces of loaders and models

Qualitative Results

Left to Right: my model, authors' model, 128x128 model, scene graph



- Models tend to capture main idea
  - "Blurry" feel, can be much worse
- My zebra seems to be floating ☺
- Their high-res model (in paper), displays two sheep, low-res does not
  - Ours better captures two sheep in last row
- Overall my model performs comparably, though edge to their model
  - Lack of training details not surprising

Visual Question Answering Results

Accuracy	Ground truth image	My model	Authors' model	# Questions
One Word	0.373	0.358	0.362	2031
Two Words	0.182	0.281	0.289	121
Three Words	0.062	0.00	0.01	97
Total	0.349	0.338	0.343	2249

Performances are low and similar!

1. Many questions are very difficult
2. Many questions are very easy
3. Memorization and contextual clues mean answerer can "cheat"

Edge to their model

- Question Answering differences can point to differences in image quality

Visual Question Answering Results

Left to Right: ground truth, my model, authors' model, scene graph

1. Many questions are very difficult
   - "What is in the hand of boy on the bench with the hat on?" (all missed: sandwich)
2. Many questions are very easy
   - "What color are the trees leaves?" (all correct: green)
3. Memorization and contextual clues mean answerer can "cheat"
   - "What color is the building?" (all correct: brown) – not even in scene graph!

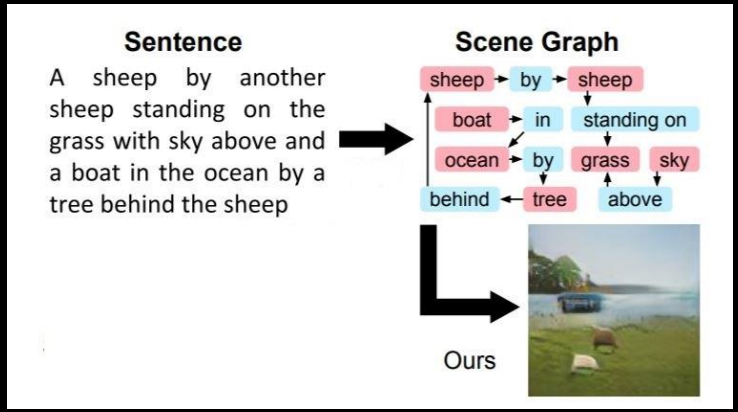
Visual Question Answering Results

Left to Right: ground truth, my model, authors' model, scene graph

- Question Answering differences can point to differences in image quality
  - "What covers the ground" (ground truth and authors' correct: snow. Mine incorrect)

Conclusion

- Task is exciting but difficult, and difficult to evaluate
- I was able to get close to performance qualitatively and using question answering metric
- Question answering metric has drawbacks but helped compare models, give deeper understanding of model and dataset



[1] Image credit: Johnson et al., *Image generation from scene graphs*. In CVPR 2018.