An Enhanced Sketch2Scene Using Natural Language

Suhong Kim, Vishun Sanjay Ramiya Srinivasan, Sara Jalili
Department of Computing Science, Simon Fraser University

Introduction

Translating a sketch into a scene is a challenging computer vision task. Realistic and meaningful scenes should contain multiple objects as well as a corresponding background, which is hard to extrapolate from a single input object. In this project, we propose a new approach to enhance the Sketch2Scene translation task using natural language processing techniques. Below is our scope of the project.

- implement a new pipeline to generate realistic scenes from a single sketch
- show how NLP can improve the Sketch2Scene task with respect to quality and diversity

Methods

The pipeline has three steps. First, Sketch Classifier predicts the class name of the object in the input sketch. Then, Caption Generator produces diverse captions based on the estimated most-similar words of the class name using retrofitted word2vec embeddings. Finally, the Caption2Scene model retrieves the objects and composes them to generate a new scene.

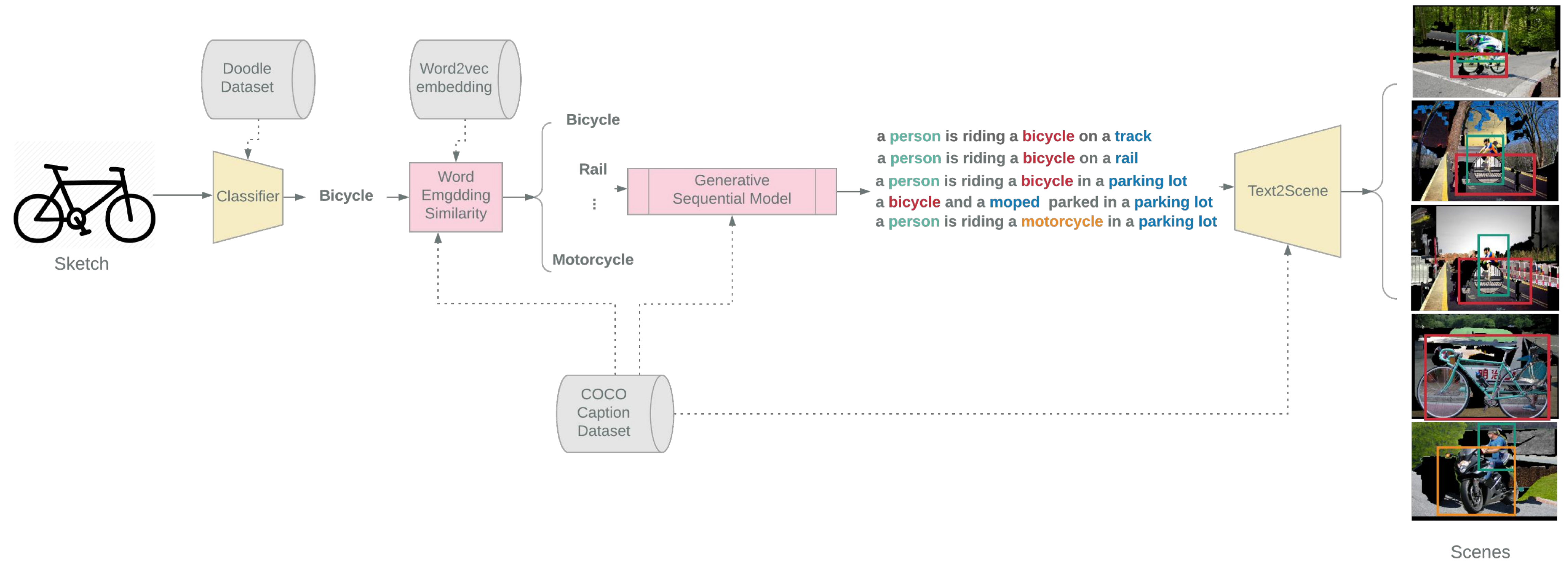
Sketch Classifier : MobilenetV2 pre-trained on ImageNet is fine-tuned using google Quickdraw dataset .

Caption Generator :

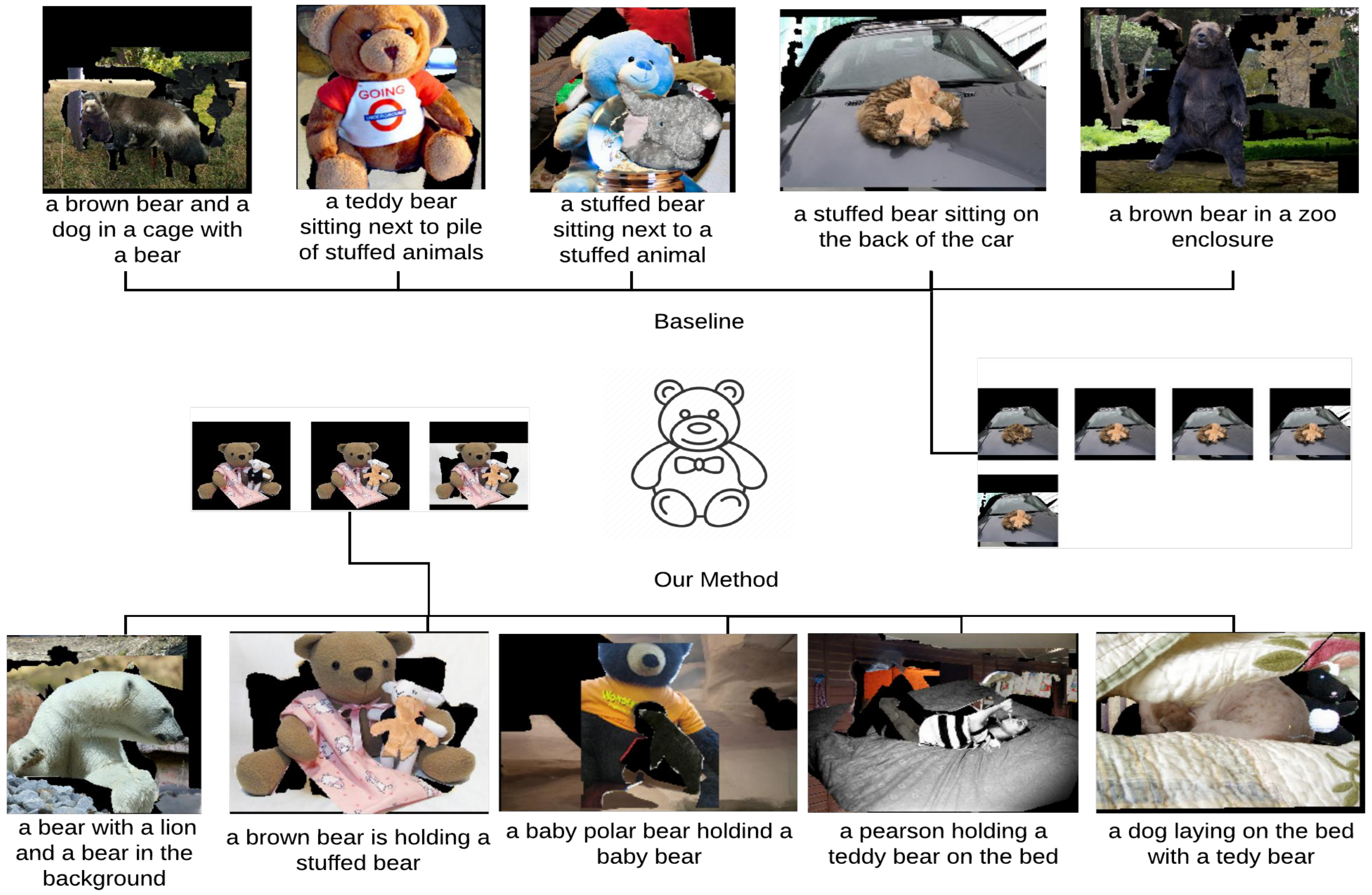
- **Baseline**: Words similar to doodle class output from word2vec embeddings are used as a condition to generate captions.
- **Our Method**- We retrofit the GloVe embeddings using object-relationship of the COCO caption dataset to improve the similarity among multiple objects in a scene [1]. During training, we randomly sample conditional words from top-10 similarity, and for inference, use beam search to improve the quality and diversity of the generated captions.

Caption2Scene : We infer the scenes using a pre-trained text2scene [2] model which learns to retrieve objects and arrange them in the scene using the semantic relationship of the objects in the captions.

Pipeline


Doodle Dataset
Word2vec embedding
Sketch
Classifier
Bicycle
Word Emgdding Similarity
Bicycle
Rail
⋮
Motorcycle
Generative Sequential Model
a person is riding a bicycle on a track
a person is riding a bicycle on a rail
a person is riding a bicycle in a parking lot
a bicycle and a moped parked in a parking lot
a person is riding a motorcycle in a parking lot
Text2Scene
COCO Caption Dataset
Scenes

Results


a brown bear and a dog in a cage with a bear
a teddy bear sitting next to pile of stuffed animals
a stuffed bear sitting next to a stuffed animal
a stuffed bear sitting on the back of the car
a brown bear in a zoo enclosure
Baseline
Our Method
a bear with a lion and a bear in the background
a brown bear is holding a stuffed bear
a baby polar bear holdind a baby bear
a pearson holding a teddy bear on the bed
a dog laying on the bed with a tedy bear

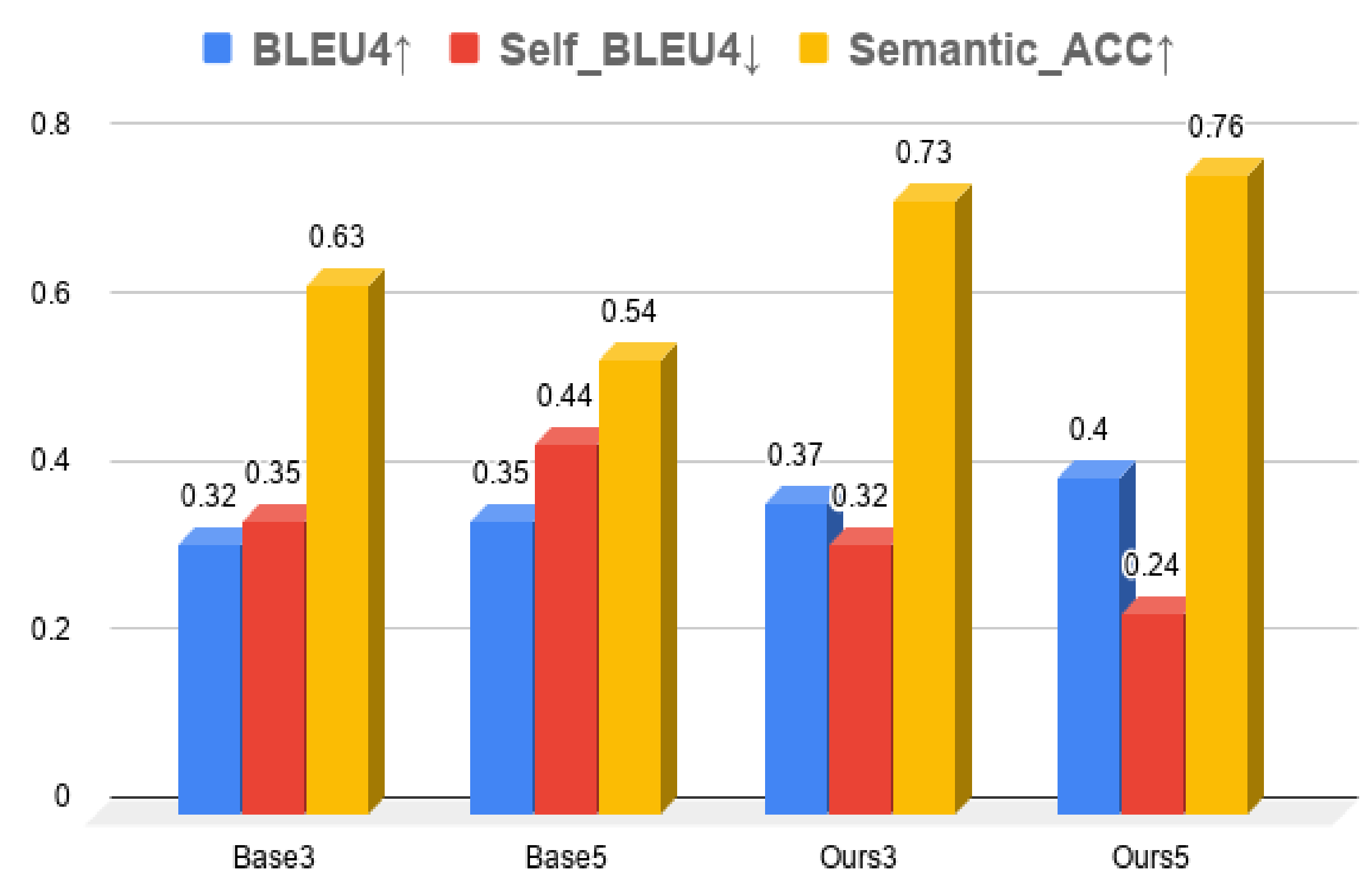
References

[1] Faruqui et al. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

[2] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating abstract scenes from textual descriptions. *arXiv preprint arXiv:1809.01110*, 2018.

Evaluation and Discussion

Caption Generator



Our method produces high quality (**High BLEU4 score**) and more diverse (**Lower Self-BLEU4 score**) captions given sketch input. Our method generates captions which have better correspondence with input sketch (**Higher Semantic Accuracy**)

Caption2Scene Generator		
Method	Inception Score	Semantic Object Accuracy
Base3	47.76	**48.3** %
Ours3	**49.10**	22.5 %

Our method generates better captions resulting in better scenes (**higher Inception score**). Caption2Scene model generates the scenes with more diverse objects and relationships, which are unseen by YOLOv3 detector (**Lower SOA score**)

Future work

- We can extend this model to produce additional data to improve object detection and image captioning task
- The end-to-end architecture will help to produce more corresponding scenes with input sketch with respect to pose and style of objects
- Conditional GAN can replace our conditional sequence generator to improve the diversity of captions