Verifying Spatial Relationships in the Visual Genome

Miroslav Vitkov

May 9, 2019

1 Introduction

The purpose of this investigation is to determine the reliability of relative spatial relationships in the Visual Genome dataset. The software design is as follows. Firstly, region descriptions are parsed to fit ¡subject phrase¿ ¡action verb¿ ¡spatial preposition¿ ¡object¿. Then rules are manufactured for each case, expressing the relationship between bounding box positions. The rules are evaluated and if false, the image, relationship and bounding boxes of the subject and object are displayed for verification by a human operator.

2 Purpose

The Visual Genome is an impressive dataset used currently and in the future. It was generated using a large number of human workers. Software heuristics were deployed to guarantee the high quality of the data(diversity in data and questions, melding of bounding boxes etc.).[**krishna**]

Nevertheless, due to the sheer amount of data points, I expect to find some errors in the annotations.

3 Parsing Relationships

Space and time are deeply ingrained in the human psychology.[1] This results in severe difficulties in differentiating actual spatial expressions from figurative ones. For example(image id = 1) "the clock is green in colour" or "a bike in the distance". Differentiating between verbatim and figurative spatial relationships is a research topic in itself.

4 Generating Rules

Once verbatim spatial relationships have been detected, rules need to be devised. Those rules are to express all legal relative positions of the bounding boxes of

the subject and object.

This task is a perfect candidate for unsupervised learning. Armed with a spatial relationship detector and the coordinates of the bounding boxes, it should be easy to detect outliers.

A less sophisticated approach is to hand craft the rules. This is facilitate by the zipfian distribution of relationship frequency, many of the top relationships being spatial.[**krishna**] For example, "on" can be modelled as "the bottom of the object's bounding box must be above the bottom of the subject's bounding box". Or "in" - "the subject's bounding box must be completely encompassed by the object's bounding box".

5 Human Verification

The purpose of human verification is twofold. Firstly, it is intended to ensure no false positives occur. The secondary goal is for the human operator to observe failure modes of the violation detection software and it's rules to facilitate improving it.

6 Details on the Software

The Visual Genome API tutorial was written for python2 - a legacy language. The associated source with this paper features the same code ported to python3. The amount of code is small, but some of the changes are non trivial.

7 Conclusion

The provided time frame limited the technical work to providing only a partial implementation of the ideas expressed above. Nevertheless, the written source code is a stepping stone to using the Visual Genome with python3 over the network or with a downloaded dataset.

References

[1] Jordan Zlatev. "Spatial Semantics". In: ().