READING ASSIGNMENT 3

16785-Integrated intelligence in robotics

[redacted]

1. What are you talking about? Text-to-Image Coreference

This paper is fairly primitive and is one of the first sources of text to image coreferencing. In this paper, Kong et. al aims to improve semantic perception by using natural language descriptions of RGB-D scenes. It is successful is 3d object detection and determining which object in the image each noun/pronoun refers to. It also supersedes expectations by achieving a deeper understanding of the scenes and solving the text to image alignment problem as well as the coreference resolution problem.

The model deployed is the Markov Random Field model which is trained on the challenging NYU-RGBD v2 dataset. Fairly decent results are obtained. When compared with Stanford conference system, it outperformed. It was also the first of its kind to improve semantic perception by integrating natural language processing and visual aids.

It can be improved by classifying aspects of the image according to verb and adjectives instead of just nouns. Furthermore, sentences could be broken down into phrase modules contained more than a single word for greater contextual reasoning. Also the model could have been trained on multiple datasets to ensure better generalization.

Another area of improvement could be the natural language descriptions of a scenes. Multiple people could be asked to describe each scene to account for variations in description in the spoken language. Also a scene could be evaluated in terms of questions such as what color and which location. Besides, recurrent neural networks or convolutional neural networks can be used as models for better text to image alignment.

2. Densely Connected Convolutional Networks

This paper has introduced a new version of convolutional networks architecture known as DenseNet. Traditional convolutional networks with n layers have n connections. However this network deploys n(n+1)/2 connections. There are direct connections between 2 layers of the same size. The features of all preceding layers and the current layer serve as inputs for the subsequent layer.

The advantages of this architecture range from better feature propagation. feature reuse, fewer parameters and overcoming the vanishing gradient issue. The proposed DenseNet is evaluated on four state of the art datasets: CIFAR-10,CIFAR-100, SVHN, and ImageNet. Overfitting is avoided and it is more efficient computationally.

However compared to algorithms like RESnet, DenseNet uses a lot more memory. The DenseNet architecture if not properly handled, normalization and contiguous convolution operations can produce feature maps that grow leaps as network depth increases. Strategies must be introduced to reduce memory consumption. This architecture can also be extended to several other computer vision tasks since they have yielded promising results so far.

In the future, capsule networks could be refined and could also be trained to outperform DenseNet on several computer vision and image perception tasks.