Kruti Thukral
012586041

Individual Project

"Text Generation from Image"

Title

Text Generation from Image

Description

Vision is an important sense for human perception and language is essential for communication between humans. With advancement in technology and robotics, building a system that can concurrently process visual stimuli and describe them in natural language would be a powerful tool to have [1]. Such a system can be used in chatty robots in wide application areas. For example, in the futuristic world, personal chatty robots can assist blind and disabled people with their day-to-day activities. Such robots are aware of their surroundings and provide continuous feedback. They act as the primary caregiver in situations wherein human caregiver is not available. Thus the building of such an intelligent system has substantial positive social impact.

Dataset

We will need a corpus of images to train the model. Some of the datasets that could be used are as below:

- Microsoft COCO - large-scale dataset for object detection and captioning [2]
- Visual Genome - Dataset to connect well-defined image concepts to language [3]
- Flickr8k
- Flickr30K

Methodology

Due to advancement in AI and machine learning, a potential solution can be implemented using deep learning. There are various alternatives to text generation. For example, an image can be described by a single high-level sentence [4, 5, 6, 7]. With such an approach, there is a restriction on the conveyed information. One recent alternative to single sentence generation is dense captioning which identifies regions of an image and then describes each with a short phrase [8]. This approach has the potential to convey more information however content need not be necessarily coherent. To overcome this limitation, a hierarchical model has been suggested that can leverage the compositional structure of both images and language [1]. Such systems to generate natural language based on an image consist of two main modules: Image Processing Module and Text Generator Module. The basic building blocks are a combination of CNNs and RNNs. CNN is used to extract image features and an RNN is used to describe image into a sentence. As part of this project, research on above-mentioned approaches will be carried out. As per feasibility, one of the approaches will be implemented and evaluated.

Evaluation

Following metrics can be used to evaluate performance

- Run time performance
- BLEU(Bilingual Evaluation Understudy)

References

[1] Jonathan Krause, Justin Johnson, Ranjay Krishna, Li Fei-Fei:
A Hierarchical Approach for Generating Descriptive Image Paragraphs,
In IEEE CVPR conference, 2017
[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan,
P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common
objects in context. In ECCV, 2014.
[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz,
S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein,
and L. Fei-Fei. Visual genome: Connecting language and
vision using crowdsourced dense image annotations. arXiv
preprint arXiv:1602.07332, 2016.
[4] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent
visual representation for image caption generation. In CVPR,
2015.
[5] J. Donahue, L. Anne Hendricks, S. Guadarrama,
M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell.
Long-term recurrent convolutional networks for visual
recognition and description. In CVPR, 2015.
[6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments
for generating image descriptions. In CVPR, 2015.
[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and
tell: A neural image caption generator. In CVPR, 2015.
[8] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully
convolutional localization networks for dense captioning. In
CVPR, 2016.