TPA 12: Image Captioning

January 26, 2017

Problem Statement: The problem introduces a captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The image captioning task generalizes object detection when the descriptions consist of a single word. Given a set of images and prior knowledge about the content find the correct semantic label for the entire image(s).

Input: An image.

Expected Output: Natural language description of the input image.

Dataset:

- **MS COCO:** The dataset contains photos of 91 object types, with a total of 2.5 million labeled instances in 328k images.
- **Flickr30k:** This dataset consists of 30K images and 150K descriptive captions.

References

- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European Conference on Computer Vision (pp. 740-755).
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. IEEE Conference on Computer Vision and Pattern Recognition (pp. 3156-3164).
- Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. IEEE Conference on Computer Vision and Pattern Recognition (pp. 4565-4574).
- Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2016). A Hierarchical Approach for Generating Descriptive Image Paragraphs. arXiv preprint arXiv:1611.06607.

- Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollr, P., & Lawrence Zitnick, C. (2015). From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1473-1482).