Image Captioning

Computer Vision (CS6350)
TPA - 4

1 Problem Statement

The problem introduces the captioning task, which requires a computer vision system to both localize and describe salient regions of images in natural language. Given a set of images and prior knowledge about the content, the correct description for the entire image(s) needs to be generated.

2 Input

- An image

3 Expected Output

- Natural language description of the input image.
- Demo to run on a given image.

4 Dataset

- **MS-COCO**: The dataset contains about 123K images covering 80 object types. Five written caption descriptions are provided for each image.
- **Flickr30k**: This dataset consists of 30K images and 150K descriptive captions.

  The evaluation can be done following the 'Karpathy' splits [1] which has been used extensively for reporting results in the literature.

5 References

1. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3128-3137).
2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

3. 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).

4. Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., & Salakhutdinov, R. R. (2016). Review networks for caption generation. In Advances in Neural Information Processing Systems (pp. 2361-2369).

5. You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4651-4659).

6. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 1, No. 2, p. 3).

7. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 3, No. 5, p. 6).