Visual Storytelling: AI-Powered Image Caption Generation

Shruti Avinash Ghorpade (22204706) and Aditya Pratap Singh (22202180)

INTRODUCTION

Step into a world where pictures transform into stories, and machines learn to speak the language of images. In today's digital age, the convergence of artificial intelligence and computer vision has propelled the development of novel applications that bridge the gap between visual data and human language. Among these, image captioning stands out as a compelling endeavor, wherein machines are tasked with automatically generating descriptive textual explanations for images. This transformative technology holds the potential to revolutionize various domains, including accessibility for visually impaired individuals, content organization, and immersive human-computer interactions, thereby motivating extensive research and innovation in the field of AI-driven image captioning.

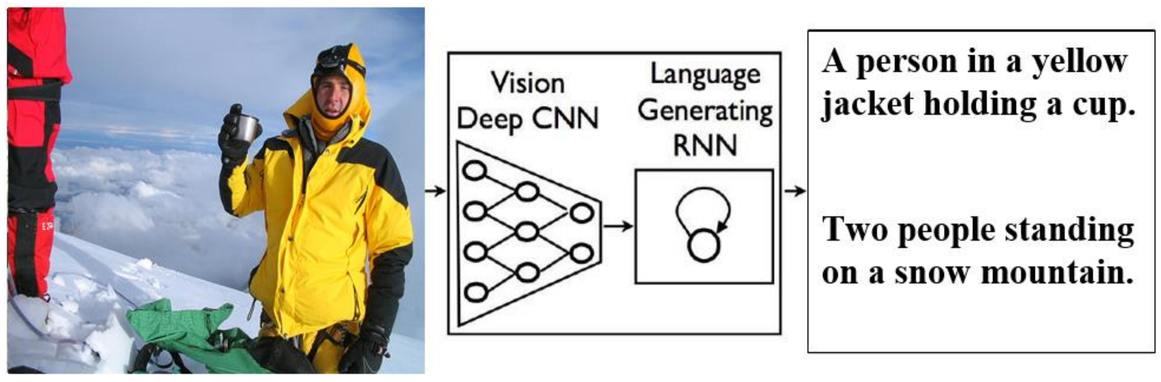


Figure: The system employs a neural network with a vision CNN and a language generating RNN, crafting coherent sentences from input images, as demonstrated in the above example.

MOTIVATION

Capturing the spirit of innovation, our motivation stems from the untapped potential within the realm of image captioning. We are driven by the aspiration to amplify the interpretive capabilities of AI, allowing it to not only recognize but deeply understand visual scenes. The potential of AI-driven image captioning across diverse sectors, enabling machines to describe images through text and bridging gaps for the visually impaired. It enhances content management, user experience in e-commerce, education, and entertainment, advancing our understanding of images and language, driving innovative approaches for accurate, coherent, and contextually relevant captions. As human-machine comprehension of visuals merges, the urge to refine AI-powered image captioning intensifies.

DATASET

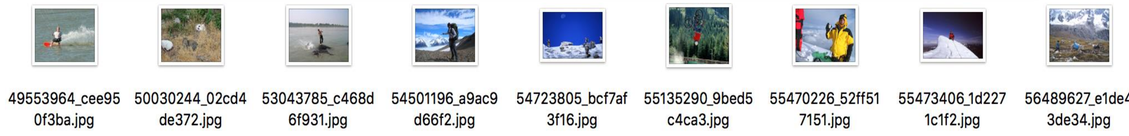
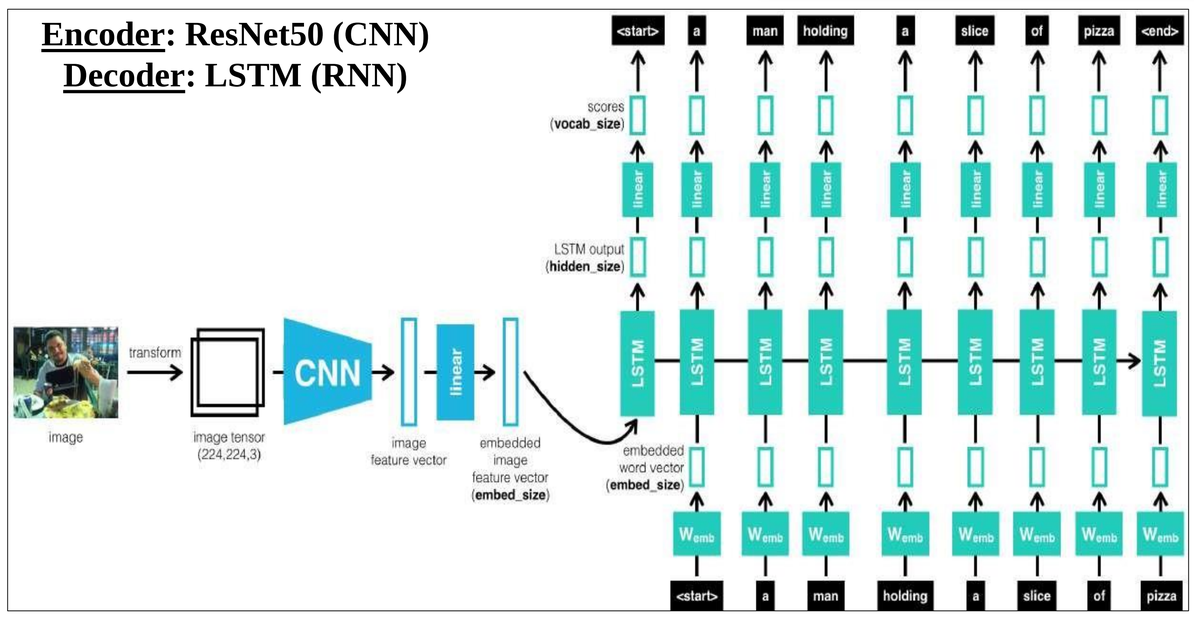


Figure: Snippet of the Flickr8k dataset from Flickr

Size and Content: The Flickr dataset comprises 8,000 high-quality images, each accompanied by five human-generated captions, resulting in around 40,000 captions. The images cover diverse scenes and activities, making it suitable for training and evaluating image captioning models.

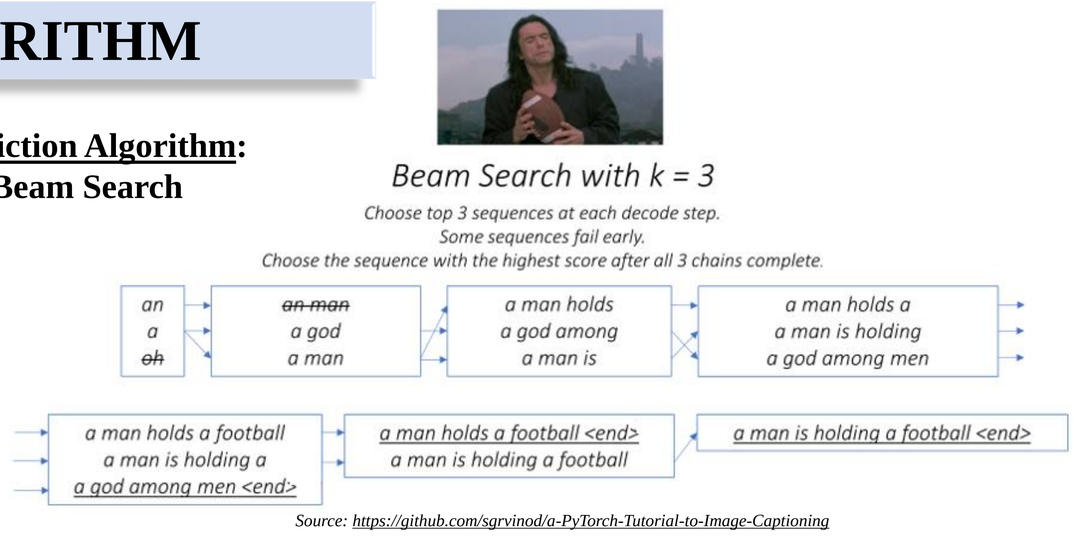
Annotations: Multiple captions per image provide varied perspectives and linguistic nuances for training image captioning models.

MODEL ALGORITHM



Source: https://raw.githubusercontent.com/sauravraghuvanshi/Udacity-Computer-Vision-Nanodegree-Program/aa9bf1ee6ed4c22477bb89d3933899ab8bca6dc3/project_2_image_captioning_project/images/encoder-decoder.png

Prediction Algorithm:
Beam Search

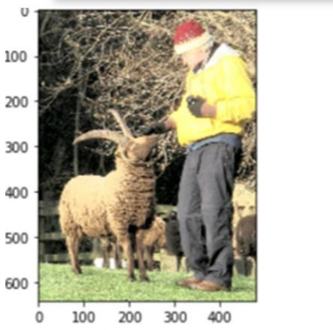


Source: https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

Source: https://github.com/ttengwang/Caption-Anything

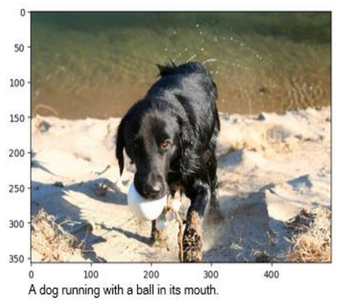
Image depicting the attention to salient features in a cluttered image

RESULTS



A man and a dog are standing next to a cow.

BLEU Score: 0.3825193716



A dog running with a ball in its mouth.

BLEU Score: 0.8012631517

A group of women standing in people.

BLEU Score: 0.5964232678

CONCLUSION AND FUTURE SCOPE

Here, we have an end-to-end neural network which not only detects and classifies images but also generates the most relevant captions in a natural language. A CNN is successfully able to encode an image into a compact representation, further a recurrent neural network is able to generate a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image. BLEU metrics is incorporated to evaluate the quality (relevance) of generated sentences.

There is a large scope of advancement to this project, for instance it can be incorporated for multimodal captioning involving modalities such as audio and video. The speed of caption generation models can be improved to enable real-time captioning for live events, videos, and other dynamic content. Further we can train Multi-linguistic models which can describe images in multiple languages. This model can be fed to a text-speech model for visually impaired or can be converted to Braille language as well.

We hope that the results of this project will encourage future work in using visual attention.

REFERENCES

1. D. Donahue, J. Jeff, L. Anne Hendrikcs, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," arXiv:1411.4389v2, Nov. 2014.
2. D. Elliott and F. Keller, "Image description using visual dependency representations," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013.
3. H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, et al., "From captions to visual concepts and back," arXiv:1411.4952, Nov. 2014.
4. M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," Journal of Artificial Intelligence Research, pp. 853-899, 2013.
   A. Karpathy and F. Fei-Fei Li, "Deep visual-semantic alignments for generating image descriptions," arXiv:1412.2306, Dec. 2014.
5. R. Kiros, R. Salahutdinov, and R. Zemel, "Multimodal neural language models," in International Conference on Machine Learning (ICML), pp. 595-603, 2014.
6. R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," arXiv:1411.2539, Nov. 2014.
   A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NIPS), 2012.
7. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 35, no. 12, pp. 2891-2903, 2013.