Comparing Image Captioning Models: Using VGG16 and DenseNet201

Srilekha Rayedi, Harshith Makkapati
Department of Computer Science
University of Houston
srayedi@cougarnet.uh.edu, hmakkakpa@cougarnet.uh.edu

Abstract

The project compares the performance of two advanced deep learning models, VGG16 and DenseNet201, in generating image captions using the Flickr8k dataset. The system capitalizes on the feature extraction capabilities of Convolutional Neural Networks and Long Short-Term Memory (LSTM) networks that generate language. We will look upon both the models to identify the computational efficiency trade-off with the quality of generated captions.

1 Research Problem

Image captioning is a crucial task in today's multimedia age, supporting accessibility, search engine optimization, and media management. However, it poses challenges as it requires understanding visual scenes, recognizing objects and their relationships, and generating coherent text. This paper addresses the need for robust captioning models by integrating CNNs for feature extraction and LSTMs for text generation. We compare two CNN architectures, VGG16 and DenseNet201, using pre-trained weights to capture visual features, paired with LSTMs for accurate captioning. Our focus is on analyzing the performance differences between these models to improve the efficiency and accuracy of image captioning.

2 Challenges

Challenges in Image Caption Generation: Major challenges tied to image captioning include:

2.1. Feature Extraction Complexity: Choosing the appropriate CNN architecture is a challenging task. VGG16 has a very simple structure and may fail for complex images that require finer detail, while DenseNet201 improves feature reuse at the cost of more computation, which makes real-time image captioning a balance between accuracy and efficiency.

2.2.Training and Computational Costs: Training and Computational Costs: Training deep learning models on large datasets is computationally expensive, despite the fact that we will be using the smaller dataset, Flickr8k, to reduce training time, high-capacity models still require huge resources. Memory management, hyperparameter tuning, and learning rate schedulers are optimization techniques that will be highly essential in preventing overfitting and ensuring stable model performances.

2.3. Caption Quality and Evaluation: generation of the meaningful and context-aware caption is a challenging task. The LSTMs have potential issues with the long-term dependencies which make the generated captions incoherent. And it is even tougher to evaluate the captions; BLEU score does not completely represent semantic accuracy. This again raises the requirement of multiple evaluation metrics necessary for quality and efficiency assessment.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

3 Potential Solutions

3.1. Feature Extraction using CNNs: VGG16 and DenseNet201, pre-trained on large datasets, are used in this paper to extract high-level visual features that provide input for the LSTM network in order to generate the captions.

3.2. Caption Generation with LSTMs: The extracted features are fed into an LSTM network in order to produce coherent captions. LSTMs will be used because of their superior context preservation capability in longer sequences.

3.3. Dataset and Preprocessing: The Flickr8k dataset will be used, comprising 8,000 images, each with five captions. The data would then be divided into training, validation, and test sets. Captions preprocessing would involve punctuation removal, stop words, and special characters, followed by tokenization, and the addition of start and end tokens.

3.4. Evaluation Metrics: The quality of the caption will be judged on BLEU score and computational efficiency by comparing the model w.r.t. training time, inference time, and memory usage.

4 Review of Related Works

Before deep learning, image captioning was based on manually extracted features combined with template-based systems. Most of these captions were shallow and lacked much context and generalization on unseen images. Neural networks automated feature extraction in this process, hence enabling the generation of more meaningful text. Farhadi et al. combined visual and linguistic information in a model that paved the way for the end-to-end model of Vinyals et al., which combined CNNs for feature extraction and LSTMs for coherent caption generation. This yielded very human-like captions, in particular when large datasets such as MS-COCO were utilized.

Advanced CNN architectures have further enhanced the captioning systems. VGG16 is one of the deep models with a large number of convolutional layers, very supreme in extracting complex features and finding wide applications in vision tasks. DenseNet201 introduces dense connectivity as a way of feature reuse and improves the flow of gradients and maintains high accuracy using relatively fewer parameters. Due to their different strengths, a comparison of the various architectures on the captioning task provides insight into how to balance computational efficiency with caption quality.

5 Conclusion

The present work tested VGG16 and DenseNet201 CNNs coupled with LSTMs for image captioning on the Flickr8k dataset. Thus, it strikes a balance between high computational efficiency and quality of captions. It is assumed that the model with VGG16 will be faster, while DenseNet201 can propose better captions. This therefore means that the results are likely to guide model selection in real-world applications by capturing some challenges perceived in providing coherent context, coupled with computational overhead. Future work can further be extended by exploring larger datasets and Transformer-based models for further improvement.

References

[1] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D. (2010). Every Picture Tells a Story: Generating Sentences from Images. Proceedings of the European Conference on Computer Vision (ECCV), 15-29.

[2] Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3156-3164.

[3] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4700-4708.

[4] Simonyan, K., Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.