Pictophrases - Teaching computers to describe images.

Darshan Rajopadhye, Kevin Heleodoro, Poornima Jaykumar Dharamdasani, Saumya Gupta
Khoury College of Computer Sciences, Northeastern University

Abstract : This study explores the application of various image captioning techniques, including classic Recurrent Neural Networks (RNNs), Bidirectional LSTMs, and Transformers, in conjunction with the InceptionV3 model, to evaluate their performance on the Flickr-8k dataset comprising approximately 8000 diverse images. Image captioning models have proven valuable in enhancing accessibility, content indexing, and retrieval across domains such as e-commerce and healthcare. However, existing models often struggle to generalize to out-of-domain images with novel scenes or objects, lacking a comprehensive understanding of context, relationships, and fine-grained details. By comparing the performance of different techniques, this comparitive study aims to identify the most effective approach for generating captions for complex images, addressing the need for improved generalization and contextual understanding in image captioning models.

1. Introduction

Image captioning is the task of generating relevant captions for the images that define the objects and actions being performed in the image. The goal is to enable machines to understand and articulate the content of images in natural language.Current image captioning models might struggle with understanding complex contextual relationships within images, leading to occasional inaccuracies. Models trained on a specific domain may not generalize well to diverse sets of images, impacting their applicability across different contexts.Models trained on biased datasets may generate captions that reflect and perpetuate societal biases present in the training data.

Our aim is to study the different models that are available and are generally used for this task of image captioning. Conducting a comparative study and understanding different methods of image captioning serves several valuable purposes in the field of artificial intelligence and computer vision. A comparative study of image captioning methods is essential for guiding the selection of appropriate models for specific tasks, fostering innovation, and advancing the overall understanding of how different architectures perform in diverse scenarios.

We have employed deep Convolutional Neural Networks- InceptionNetV3 for image feature extraction and have used Recurrent Neural Networks, Bidirectional Long Short-Term Memory (LSTM), and self attention based Transformers for generating text captions. The models will be assessed using the BLEU evaluation metrics to ensure the quality and accuracy of generated captions. Here, we are trying to understand the architectures of these different models, how they are able to capture the details and contexts of images and what are their advantages and limitations.

2. Related Work

In this section, we provide an overview of relevant studies in the field of image captioning, highlighting key advancements, methodologies, and contributions. The exploration of related work aims to situate our research within the broader landscape, offering insights into the evolution of image captioning techniques and the challenges addressed by previous studies.

2.1. Deep Learning Approaches on Image Captioning: A Review

Ghandi and Pourreza [1] conducted a comprehensive review, highlighting the transformative impact of deep learning and vision-language pre-training on image captioning evolution. They categorized deep learning methods, including R-CNNS, RNNs, LSTMs, GRUs, and Transformers, and extensively explored attention mechanisms, covering soft and hard attention. The survey addressed challenges such as contextual under-

standing, object hallucination, and missing context, providing valuable insights by ranking deep learning methods and suggesting future research directions.

2.2. Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning.

An end-to-end trainable deep bidirectional LSTM (Bi-LSTM) model presented by [2] integrates a deep CNN and two separate LSTM networks to capture long-term visual-language interactions effectively. The model incorporates data augmentation techniques to prevent overfitting during training, demonstrating competitive performance in caption generation and image-sentence retrieval across benchmark datasets. The research underscores the benefits of multi-task learning for model generality and the effectiveness of transfer learning with the Bi-LSTM model.

2.3. Transformer-based Local-Global Guidance for Image Captioning

The authors introduced a transformer-based model with a focus on local-global guidance for image captioning [3]. This novel approach leverages the strengths of transformers, specifically their attention mechanisms, to enhance both local and global context understanding. By incorporating local-global guidance, the model aims to capture detailed information within images while maintaining a broader context. The transformer-based architecture efficiently processes image features, enabling the generation of coherent and contextually rich captions. This work contributes to the evolving field of image captioning by introducing a novel approach that leverages transformers for improved contextual understanding, presenting a potential advancement in the generation of informative and nuanced image captions.

3. Methodology

In this study, we employed three distinct model architectures for image captioning: Gated Recurrent Unit (GRU) within a Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM), and a Transformer. Our objective was to compare these architectures on their performance in generating image captions, utilizing the Flickr-8k dataset. To keep the comparision as faair as possible we use the same, InceptionV3 image feature extraction model for all the three architectures. The selected dataset encompasses around 8000 diverse images, making it suitable for evaluating the models' capabilities in handling varied scenes, objects, and concepts. For a standardized assessment, we employed the BLEU metric, a widely accepted measure for evaluating the quality of generated captions in comparison to reference captions.

3.1. Feature Extraction

We selected the InceptionV3 [5] model as our feature extraction architecture for consistent comparison across different image captioning techniques. InceptionV3 is renowned for its effectiveness in capturing intricate hierarchical features from images, offering a balance between model complexity and computational efficiency. Its ability to detect multi-scale visual patterns makes it suitable for a wide range of scenes and objects present in the diverse Flickr8k dataset. By employing InceptionV3 across all our experiments, we aim to ensure a fair and uniform baseline for evaluating the diverse image captioning models, allowing us to isolate the impact of the captioning techniques themselves while maintaining a consistent and reliable feature extraction foundation. This choice contributes to the interpretability of our results, providing insights into the relative strengths and weaknesses of different image captioning methodologies while holding the feature extraction component constant.

3.2. Model Architectures

We employed three different model architectures for image captioning:

3.2.1. GRU in RNN

3.2.2. Bidirectional LSTM (BiLSTM)

3.2.3. Transformers

3.3. Dataset

We chose to utilize the Flickr8k dataset in our study as it offers a diverse and comprehensive collection of around 8000 images, each accompanied by detailed captions. The dataset was selected for its rep-

resentation of a wide array of scenes, objects, and concepts, making it well-suited for evaluating the performance of image captioning models across various real-world scenarios. By incorporating Flickr8k into our experimentation, we aimed to leverage its richness and diversity to thoroughly assess the generalization capabilities of the employed image captioning techniques. The availability of high-quality images with corresponding captions in the dataset facilitated the training, validation, and testing phases, allowing for a robust evaluation of the models' performance on a broad spectrum of visual content. Overall, the choice of the Flickr8k dataset aligns with our objective of conducting a comprehensive and realistic assessment of image captioning models in diverse and representative settings.

3.4. Metrics

The authors introduced BLEU [4], a method for automatic evaluation of machine translation, in their paper published in 2002. BLEU (Bilingual Evaluation Understudy) has since become a widely adopted metric for assessing the quality of machine-generated translations by comparing them to human-generated reference translations. This metric measures the overlap of n-grams between the machine-generated and reference translations, providing a quantitative evaluation that aligns with human judgments of translation quality.

4. Results & Discussion

In some cases, it might be prefered to split this part into two sections.

5. Conclusion

Although a conclusion may review the main points of the paper, it must not replicate the abstract. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Do not cite references in the conclusion. Note that the conclusion section is the last section of the paper to be numbered. The appendix (if present), other acknowledgments, and references are listed without numbers.

Acknowledgments

Use this section to thank anyone besides the main author(s) who has contributed to the paper. Avoid expressions such as "One of us (S.B.A.) would like to thank. . . " Instead, write "F. A. Author thanks. . . ".

References

[1] Ghandi, T., Pourreza, H., and Mahyar, H., "Deep Learning Approaches on Image Captioning: A Review," *ACM Comput. Surv.*, Vol. 56, No. 3, 2023. doi: 10.1145/3617592, URL https://doi.org/10.1145/3617592.

[2] Wang, C., Yang, H., and Meinel, C., "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, Vol. 14, No. 2s, 2018. doi: 10.1145/3115432, URL https://doi.org/10.1145/3115432.

[3] Parvin, H., Naghsh-Nilchi, A. R., and Mohammadi, H. M., "Transformer-based local-global guidance for image captioning," *Expert Systems with Applications*, Vol. 223, 2023, p. 119774. doi: https://doi.org/10.1016/j.eswa.2023.119774, URL https://www.sciencedirect.com/science/article/pii/S0957417423002750.

[4] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J., "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, Vol. 2, 2014, pp. 67–78.

[5] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the Inception Architecture for Computer Vision," , 2015.

[1–5]