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. Dataset

We chose to utilize the Flickr8k dataset [4] 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 representation 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 while still being computationally efficient for the scope of the project as compared to other datasets in practice.

3.2. 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.

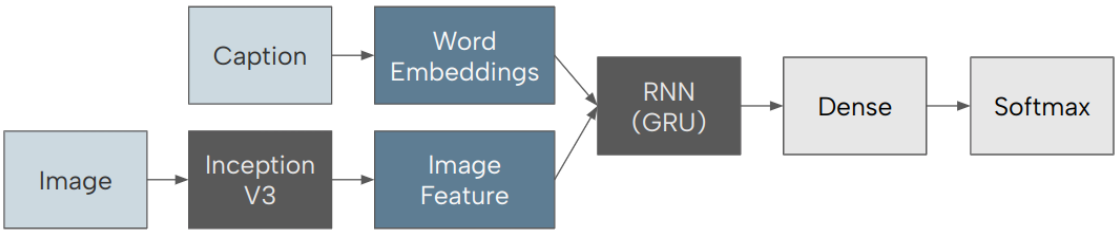

Figure 1. RNN Model

3.3. Model Architectures

We employed three different model architectures for image captioning:

3.3.1. GRU in RNN

The GRU RNN model architecture for image captioning is structured to effectively combine visual and sequential information for accurate caption generation. The model takes two types of input: visual features extracted from images and sequential information from tokenized captions. The model comprises two key components as shown in figure 1. The first component processes the visual information from images, with an input layer followed by dropout regularization and a dense layer with ReLU activation. The second component handles the sequential information from the captions using an embedding layer with a trainable embedding matrix and a GRU layer. The outputs from these components are then combined and further processed through dense layers to generate a softmax output for vocabulary prediction. The model is compiled using categorical crossentropy loss and the Adam optimizer, making it suitable for training on image-caption pairs.

3.3.2. Bidirectional LSTM (BiLSTM)

Bidirectional LSTMs process inputs in two directions once forward and once backward so they capture both past and future contexts of information and are able to capture long term dependencies in data and understand the relationship between different elements in a sequence even better than LSTMs. The flow of information between each cell of LSTM is better depicted in figure 2. For the image input (images having 2048 features), it applies a neural network layer to embed the image features into a lower-dimensional space using the ReLU activation function.It repeats the embedded image to match the length of captions.The resulting model takes an image as input and outputs the repeated, embedded image features. For the language input (captions), it uses an embedding layer to convert words into a numerical format. This is followed by a bidirectional LSTM layer.Dropout and batch normalization are applied to enhance the model's generalization and stability. The resulting model takes a sequence of words as input and outputs a sequence of embedded words. The outputs of the image and language models are concatenated together. This means the information from both the image and the language models are combined into a single representation. The concatenated representation goes through a bidirectional LSTM layer to capture contextual information.Finally, a dense layer with softmax activation is used to predict the next word in the sequence.The resulting model takes both an image and a sequence of words as input and outputs a probability distribution over the vocabulary for the next word in the sequence.

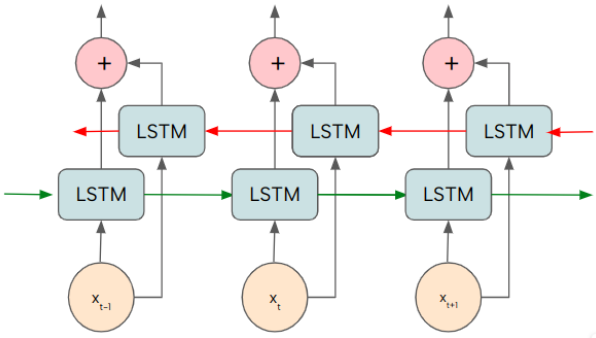

Figure 2. BiLSTM's Flow Diagram

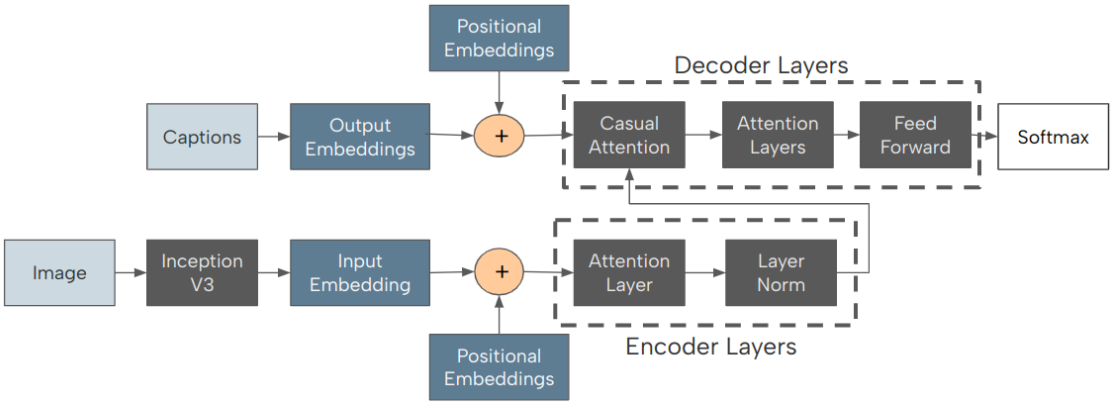

Figure 3. Transformer architecture.

3.3.3. Transformers

The Transformer models encoder-decoder architecture, as shown in figure 3, is as described below:

Encoder : The encoder's job is to process the input data—in this case, the embeddings of the image features—and generate a rich, context-aware representation. Each encoder layer consists of a self-attention mechanism followed by a feed-forward neural network. The self-attention mechanism allows the encoder to weigh the importance of different parts of the input data differently, which is particularly useful for understanding complex inputs like images where certain features may be more relevant to the captioning task than others. Layer normalization is applied after each sub-layer to help in stabilizing the activations and speeding up the training process.

Attention Mechanism : The attention mechanism allows the model to focus on different parts of the input sequence as it processes data. It helps the model to draw parallels between specific visual features and certain words in the caption. There are typically three vectors involved in the attention mechanism: the query, key, and value. The self-attention in the encoder compares all keys to each query to produce a weighted sum of value vectors. This operation gives the model its ability to pay attention to different parts of the image when predicting each word in the caption. In the decoder, the attention mechanism is used again in the form of 'causal attention.' This ensures that the prediction for the current word can only be influenced by known previous words, not by any future words in the sequence.

Decoder : The decoder takes the encoder's output and generates the prediction sequence, i.e., the caption. It also uses self-attention layers. The self-attention layers in the decoder are masked to ensure that the prediction for a given word can only depend on earlier words in the sequence. This is essential during training, where we want to avoid giving the model access to the future tokens it is supposed to predict. The decoder also includes attention layers that focus on the encoder's output, effectively allowing the decoder to use the full context of the image as provided by the encoder's output when generating each word of the caption.

Finally, the output of the decoder layers goes through a softmax layer. The softmax function is used to create a probability distribution over all possible words in the vocabulary for the next word in the caption. The word with the highest probability is chosen as the output at each step, and this process is repeated until the model generates an end token or reaches a maximum length, thus completing the caption.

3.4. Metrics

We will assess the model's performance using the BLEU [6] (Bilingual Evaluation Understudy) metric which evaluates generated text against reference text in terms of sequences of words. It first measures the precision of n-grams (contiguous sequences of n items, usually words) in the generated text compared to a set of reference texts. There is a brevity penalty to address the issue of short translations. If the generated text is significantly shorter than the reference texts, the brevity penalty reduces the BLEU score. The

individual precision scores for different n-gram sizes are combined using a weighted geometric mean to get the final BLEU score.

$$BLEU = BP + \exp(\sum_{n=1}^{N} w_n log(precision_n)) \quad (1)$$

Where:

- N is the maximum n-gram order considered (usually up to 4).
- precision_n is the precision for n-grams.
- w_n is the weight assigned to the precision of n-grams (commonly set to $1/N$, i.e., equal weight for each order).
- BP is the brevity penalty, which penalizes shorter translations to address the issue of favoring shorter sentences.

4. Results & Discussions

In our experimental results, we observed notable differences in the performance of various models for image captioning. The experimental setup and the corresponding results are discussed in these sections.

4.1. Experimental Setup

In this section, we detail the experimental setup employed to evaluate the performance of various image captioning models. The training for all the three models was either done locally on a mobile GPU or on Google Colab using a T4 GPU. The RNN and BiLSTM models were trained for 20 epochs whereas the early stopping made the Transformer model stop everytime after 17 epochs. The input dimension for all the models was 2048, the standard image feature dimension as generated by the InceptionV3 model. We used InceptionV3 with pretrained weights on imagenet to get better and consistent image features for the diverse scenes covered in the Flickr8k dataset.

4.2. Results

We present quantitative metrics to assess the performance of each image captioning model. This section provides a comprehensive analysis of how well each model performed, utilizing metric described earlier. The RNN model exhibited the least favorable outcomes, followed by the BiLSTM, while the Transformer model demonstrated comparatively superior performance. These findings are substantiated by the quality and relevance of the generated caption as showed in later sections.

Table 1. BLEU Scores for Different Models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
RNN	0.49	0.23	0.15	0.08
BiLSTM	0.43	0.24	0.16	0.08
Transformer	0.43	0.29	0.20	0.08

4.2.1. Discussion

The unexpected result of the RNN outperforming the Transformer and BiLSTM models in BLEU-1 scores while exhibiting a consistent ranking in the subsequent BLEU metrics prompts further analysis. The BLEU-1 score primarily evaluates the precision of unigrams, capturing the accuracy of individual words in generated captions. The RNN, known for its sequential processing, may excel in capturing simple, unigram-level relationships in image descriptions. However, as the n-gram order increases, the Transformer and BiLSTM, with their capacity for capturing broader contextual dependencies and long-range relationships, regain their expected positions. This anomaly highlights the nuanced nature of evaluation metrics. Because of this to get a better understanding of the actual performance of the model we include BLEU-2 and BLEU-3 in the visualization of the model performance.

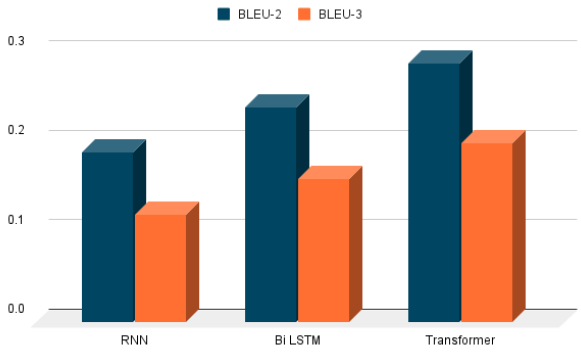

Figure 4. Metric Visualization

The RNN's limitations in capturing long-range dependencies and context appear to impact its captioning efficacy negatively. On the other hand, the BiLSTM model, with its enhanced ability to capture bidirectional context, displayed improved performance. Notably, the Transformer model, leveraging self-attention mechanisms, surpassed both RNN and BiLSTM in generating more accurate and contextually relevant image captions. These observations underscore the significance of model architecture in the image captioning task, with the Transformer architecture emerging as a promising candidate for capturing intricate relationships within images and generating high-quality captions.

4.3. Comparison

Apart from the examples displayed later we present a common example that we feel portrays the abilities of each model and also sheds some light on its limitations. In figure 5 we see an image of a person sitting on top of a rock looking down on the mountains. At first when we see the image the first clear relationship we get is that a person is sitting. Avoiding all other relationships the RNN model predicts that the man is sitting on something. It classifies the rock as a bike which is incorrect but could be because of the limited data we are working with. In the BiLSTM prediction we see that there are some dependencies that it tried to capture. We see that it tried to predict the clouds in the background as snowy mountains, which although being incorrect showcases the strengths of the model in capturing the long range dependencies that the RNN could not pick up on. Finally in thee transformer prediction we can see that it identifies the person and then also that they are on top of a rock but again the model fails to correctly identify the pose of the person which could get better if there was more training data. A key thing we noticed is that our model often gets the gender of people wrong. It's not just a one-time mistake; we saw this happening a lot in different examples, like with colors and animals too. This issue is a big challenge because it makes our model's predictions less accurate.

Upon examining true captions and model-generated descriptions, it's evident that even the best-performing model falls short of replicating the intricacies found in actual image descriptions. Addressing these disparities through subtle adjustments and alternative techniques was feasible once the initial models were designed. However, project constraints led us to prioritize specific aspects. While acknowledging the potential for implementing these improvements, we recognize that decisions were made to keep results within project scope. The Future Scope section delves into these limitations, suggesting possible avenues for refinement and advancement in image captioning models.

Figure 5. Predicted Caption Comparison

5. Conclusion

In conclusion, our exploration of different image captioning models revealed nuanced performance dynamics. While the RNN demonstrated unexpected success in BLEU-1 scores, deeper analysis unveiled its proficiency in unigram precision. However, as n-gram complexity increased, the Transformer and BiLSTM reclaimed their expected positions, emphasizing the importance of evaluating models across various metrics. It is crucial to note that while BLEU scores provide a quantitative measure of model performance, they may not fully capture the nuanced quality of generated captions, as discussed in subsequent sections. Additionally, a noteworthy observation is the marginal improvement in scores achieved by the BiLSTM compared to the RNN, considering the considerably higher computational demands of the former. This raises questions about the practical trade-off between computation cost and incremental score gains.

6. Future Scope

Looking ahead, there's room to enhance image captioning approaches. Beyond just numbers, we aim to

capture the essence of quality in captions, bridging the gap between metrics and what we see in words. It's crucial to note that BLEU scores, our metric comparison, may overlook synonyms or alternative word forms. Considering the richness of language and expression in captions, future exploration could delve into evaluation methods that better capture these nuances. The study hints that the BiLSTM might demand more power than it's worth, emphasizing the need for efficient strategies. Also, our findings suggest that the captions, while good, could be even better with more diverse data, like from MSCOCO [7], offering potential for richer and more accurate descriptions. Importantly, as we delve into the ethical dimensions of data diversity and representation in training image captioning models, it becomes evident that our dataset exhibited bias toward certain classes. While our findings underscore the necessity for a responsible and inclusive approach in dataset curation, it's crucial to note that due to time constraints, we weren't able to thoroughly investigate and address this bias. The recognition of bias in our dataset emphasizes the importance of dedicating time and effort to rectify such imbalances. Future endeavors could prioritize an in-depth analysis of dataset bias, aiming to ensure fairness and accuracy in training image captioning models. Addressing these biases is pivotal for enhancing the model's ability to generate captions that are both diverse and representative of the broad spectrum of images encountered in real-world scenarios.

7. Team Contributions

With a team of four members and three model architectures, our approach was collaborative and efficient. Kevin led data processing and feature extraction, while the remaining three each took charge of implementing and fine-tuning one model architecture. Darshan led the implementation and evaluation of the RNN, Saumya focused on the BiLSTM model, and Poornima took charge of implementing and evaluating the Transformer model.

Acknowledgments

We would like to express our sincere gratitude to Prof. Raj for his guidance and support throughout the duration of this class project. It played a pivotal role in shaping our understanding and approach to the project. We also extend our thanks to our fellow classmates for their collaborative spirit, contributing to a dynamic and enriching learning experience.

Link to Code Repository

`https://github.com/therrshan/image-captioning`

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.

[6] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, USA, 2002, p. 311–318. doi: 10.3115/1073083.1073135, URL `https://doi.org/10.3115/1073083.1073135`.

[7] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P., "Microsoft COCO: Common Objects in Context," , 2015.

[1–7]

Predicted Caption:
man is skiing down a snowy mountain

True Caption:
A skier wearing a blue jacket and helmet is skiing down a hill.

Predicted Caption:
man in red shirt is standing outside building

True Caption:
A group of people looking at sound equipment.

Predicted Caption:
man is sitting on rock by the mountain

True Caption:
A little girl balances on rocks on the beach.

Figure 6. RNN Examples

Predicted Caption:
a man in a yellow shirt is airborne on a motorcycle

True Caption:
A cyclist in a helmet is driving down a slope on his bike .

Predicted Caption:
a girl in a purple jacket is riding a blue and red and yellow coat across a field

True Caption:
A man in street racer armor is examining the tire of another racer 's motorbike .

Predicted Caption:
a baseball player in purple jersey about to kick the ball

True Caption:
A baseball player slides toward a base .

Figure 7. BiLSTM Examples

Predicted Caption:
a brown dog is running through a field

True Caption:
A brown dog is running in the grass

Predicted Caption:
a man in a blue shirt is riding a bike through the woods

True Caption:
A man wearing a blue helmet riding a bike in the woods

Predicted Caption:
a group of people are standing in front of a crowd

True Caption:
Girls in light blue outfits perform a choreographed dance

Figure 8. Transformers Examples