Machine Learning Techniques to Predict Image Captions

Arjun Pesaru
MSDS, Northeastern University
pesaru.a@northeastern.edu

Hiranmai Devarasetty
MSDS, Northeastern University
devarasetty.h@northeastern.edu

Abstract

A Image Caption Generator is a system that is built using Artificial Intelligence and Natural Language Processing to analyze an image and generate an accurate caption.This project mainly aims to bridge gap between visual content and textual description. Generating Image captions requires deeper understanding of the visual content for producing coherent and descriptive textual descriptions. Our research introduces an innovative approach for image caption generation,leveraging the different machine learning techniques such as VGG16, Inception V3 Convolutional neural networks and LSTM, an Recurrent neural network, which generates textual description for the images, one word at a time.This study investigates the effectiveness of the Image captions generated by the Machine learning techniques on the Flickr8k dataset. The exploration of VGG16, InceptionV3 and LSTM includes the data preprocessing, model building and evaluation. OPtimization of the model paramters were performed to achieve the model's optimal performance. The Flickr8k dataset is used to provide the ground-truth captions. Evaluation metrics includes BLEU scores, where the model generated captions are compared to the actual captions present in the test dataset. By bridging the semantic gap between visual content and textual descriptions, image caption generating models can enhance accessibility, improve image understanding, facilitate human-machine communication,and help visually impaired individuals. Despite promising results, opportunity for improvement exists, refining the model architecture, fine tuning the model,integrating the attention mechanisms and leveraging larger datasets. This technology has become popular and has several applications across various domains including Healthcare, Industries and disciplines.

1 Introduction

The Image Caption Generation explores the intersection of Artificial Intelligence, Computer Vision and Natural Language Processing, where the goal is to develop systems that are capable to generate descriptive and accurate captions for the images automatically. Unlike humans who can generate descriptions by just glancing at an image, task of generating captions automatically is complex, as it requires the machines to understand the visual content of an image as well the semantics of natural language to produce coherent and accurate textual descriptions.

Deep Learning Methodologies has produced great results in generation of Image Captions in recent years, the architectures like Convolutional Neural Networks(CNNs) and Recurrent Neural Networks(RNNs) has significantly advanced the capabilities of these systems by allowing them to learn hierarchical representations of visual and textual content. Convolutional Neural Networks(CNNs) can learn to detect and encode meaningful visual patterns, such as objects,Shapes,Textures, and Edges. These extracted features are served as rich presentations of image content, which are fed into subsequent layers of the network for further processing. Recurrent Neural Networks(RNNs), specifically Long-Short Term Memory(LSTM) netowrks, are powerful in sequential data modeling and are widely used for generating natural language descriptions. One of the key advantages of using deep learning is its ability to perform end-to-end learning, where both the visual features as well as the textual data are optimized jointly.The addition of attention mechanism enhances the deep learning models in image captioning, as it allows the model to focus on relevant regions on the image while generating each word of the predicted caption.

These models can be further deployed into user friendly interfaces,such as Streamlit apps, to simplify the process of generating captions for the images. By integrating the trained model into an interactive model, where users can upload the images and generate the captions without requiring the access to the underlying code. By deploying the models into such interfaces, broadens the accessibility and enables seamless use of image captioning technologies in real-world scenarios.

1.1 Review on Literature

Image caption generation is a multifaceted task that integrates Artificial Intelligence, Computer Vision, and Natural Language Processing. Recent advancements in deep learning have significantly enhanced the capabilities of image captioning systems. Convolutional Neural Networks (CNNs), such as VGG16 and InceptionV3, are widely used for extracting visual features from images, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are employed for generating sequential text descriptions.

The incorporation of attention mechanisms has been a crucial improvement in these models, enabling them to dynamically focus on relevant regions of the image during caption generation. This approach has been exemplified in models like "Show, Attend, and Tell," which have shown substantial improvements in caption accuracy. Additionally, the use of larger and more diverse datasets, such as the COCO dataset, and the exploration of Transformer-based architectures have opened up new avenues for enhancing performance. These models not only improve the accuracy of captions but also have broader implications, such as enhancing accessibility for visually impaired individuals and facilitating human-machine communication across various domains, including healthcare and industries.

1.2 Key Findings and Future Directions

Studies on image caption generation have highlighted the importance of optimizing model parameters and leveraging advanced architectures. The Flickr8k dataset, for instance, has been used to evaluate the effectiveness of CNN-LSTM models, with metrics such as BLEU scores indicating areas for improvement. Future directions include integrating multimodal learning techniques, refining model architectures with attention mechanisms, and utilizing larger datasets to capture a wider range of visual and linguistic features. The deployment of these models into user-friendly interfaces, such as Streamlit apps, further enhances their practical applicability and accessibility in real-world scenarios

2 Dataset

In this experiment, The dataset which is being used is known as Flickr8k dataset, which comprise a collection of 8000+ images with 5 captions each. The output caption for the provided image will be generated based on the trained models from the given five captions.

Table 1: DATASET

Dataset	Images	Captions for each image	Training	Testing	Validation
Flickr 8K Dataset	8091	5 captions per image: 40455 Captions in toto	60%	20%	20%

3 The Implementation

3.1 Model Architecture:

The proposed model for image caption generation is a combination of an Encoder-Decoder framework enhanced with attention mechanisms. The key components and their functionalities are described below:

1.Encoder:

1. Pre-Trained VGG16:
   - Extracts visual features from input images.
   - Outputs a dense vector representation of the image.
   - Feature vector size: 4096 (from the final dense layer of VGG16).
2. Dense Layer with Batch Normalization:
   - Projects image features into a lower-dimensional space.
   - Includes Batch Normalization to stabilize and speed up training.

2.Decoder:

1. Embedding Layer:
   - Converts input captions into dense vectors of size 64.
   - Supports masking to handle variable-length sequences.
2. LSTM Layer:
   - Processes sequential data (tokenized captions).
   - Configured with 256 units, with Dropout (30%) and Recurrent Dropout (30%) for regularization.
3. Attention Mechanism:
   - Implements an Additive Attention layer to dynamically focus on relevant parts of the image features.
   - Combines visual and textual features for better contextual understanding.
4. Dense Layers with Output:
   - Adds non-linearity and projects combined features to the output space.
   - Includes Dropout (30%) to reduce overfitting
5. Output Layer:
   - Outputs probabilities for each word in the vocabulary.
   - Uses a Softmax Activation to compute the most likely next word.

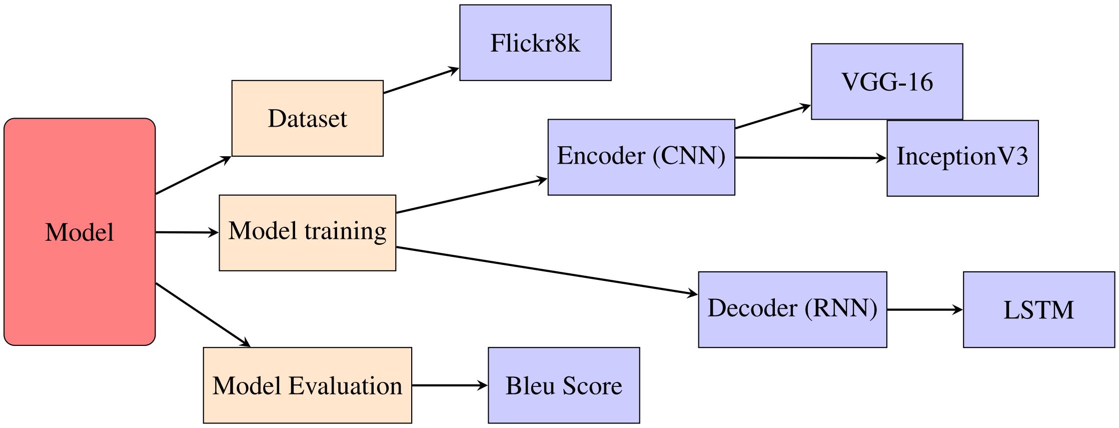


Figure 1: Workflow of the model with dataset, training, evaluation, and specific components.

As shown in Figure 1, the workflow includes dataset preparation, model training, evaluation, and detailed architecture components.

3.2 Parameters and Hyper parameters choice:

1. Model Parameters:

1. Embedding Size: 64
2. LSTM Units:256
3. Dense Layer Units: 256
4. Attention Units: Combined dynamically during generation.
5. Vocabulary Size: Matches the tokenized dataset vocabulary.

2. Hyperparameters:

1. Learning Rate: 0.0001
    - Selected for stable and fine-grained convergence.
2. Dropout: 30% on LSTM and Dense layers to reduce overfitting.
3. Regularization:
    - L2 Regularization (0.01) applied to Dense and LSTM layers.
4. Batch Size: 128
    - Chosen to balance memory efficiency and gradient stability.
5. Epochs: 20
    - With EarlyStopping to prevent overfitting when validation loss plateaus.

3.3 Training and Optimization:

1. Optimizer:
    - Adam optimizer is being used for its adaptive learning rate capabilities and computational efficiency.
2. Learning Rate Scheduling:
    - Learning Rate Scheduling dynamically reduces the learning rate when the validation loss is plateaued.
3. Early stopping:
    - Prevented unnecessary training by halting when validation loss failed to improve for three consecutive epochs.

The integration of Attention mechanism, Batch normalization and careful hyperparameter selection enabled the model to effectively learn from the Flickr8k dataset. While the dataset's limited size constrained performance, the model demonstrated robust learning and the ability to generate coherent captions. Further improvements can be achieved by scaling the dataset and exploring Transformer-based architectures.

3.4 Experimental Results:

1. Training Metrics:
    - Training Loss: Consistently decreased from 11.57 in the first epoch to 3.12 by the 20th epoch.
    - Validation Loss: Reduced from 9.59 to 4.54, demonstrating improved generalization.
2. BLEU Scores:
    - Mean BLEU Score: 0.042 across the test set.
    - Good Predictions: Captions with BLEU scores above 0.6 achieved an average BLEU of 0.68
3. Loss Progression: The loss progression for both training and validation is visualized below

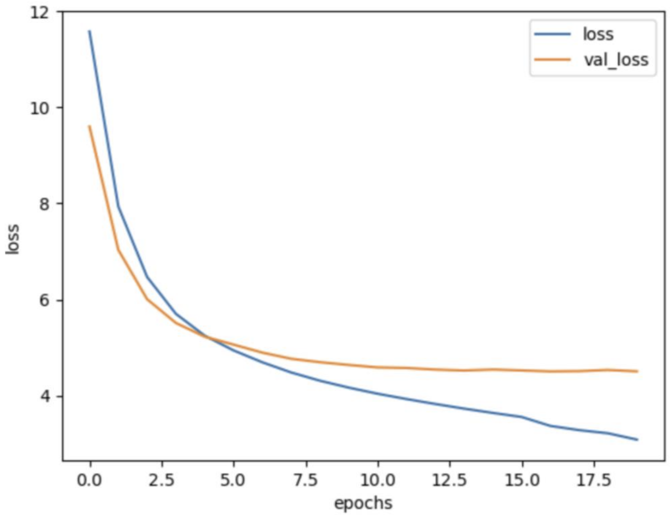


Figure 2: Loss and validation loss over epochs.

Figure 2 illustrates the training and validation loss over 20 epochs, highlighting the model's effective learning process. The training loss steadily decreases from 11.5 to around 3.0, while the validation loss stabilizes near 4.5 after the 10th epoch. This plateau indicates optimal performance on unseen data without significant overfitting. The consistent drop in training loss, alongside stable validation loss, demonstrates the model's ability to generalize effectively, though minor adjustments to regularization or dataset size could further enhance performance.

Insights:

- The model converged steadily with no significant overfitting.
- Validation loss plateaued after 15 epochs, prompting early stopping.

3.5 Examples of the Generated Captions:

1. Image:A black dog running in the grass.
   - Ground Truth: "A black dog is running in the grass."
   - Generated: "A black dog is running in grassy grass." (BLEU: 0.643)
2. Image: A family having a picnic.
   - Ground Truth: "A family is sitting on the grass having a picnic."
   - Generated: "People sitting grass food." (BLEU: 0.421)

These results suggest that the model performed well for simpler captions but struggled with more complex scenarios.

4 Results

4.1 Main Results

1. Caption Quality: The generated captions effectively described simpler images but lacked precision in complex or crowded scenes.
2. Training Dynamics: Both training and validation losses showed consistent reductions over epochs, indicating the model was learning effectively.

3. Performance Gaps: Despite the progress, the relatively low BLEU score highlights the challenge of capturing linguistic nuances, especially for diverse and complex images.

Key highlights of the Results:

- The model demonstrated its ability to generate grammatically correct and contextually relevant captions for certain scenarios.
- However, in cases involving abstract concepts or multiple objects, the captions often failed to capture the complete context, reducing BLEU scores.

Table 2: Evaluation Metrics

Name	**Values**
BLEU Score for good predictions	0.68
BLEU Score for bad predictions	0.0013

4.2 Supplementary Results

1. Parameter Choices:
   - Learning Rate: A small learning rate of 0.0001 was chosen to ensure stable and gradual learning, preventing large updates that might disrupt convergence.
   - Batch Size: A batch size of 128 was selected after experimentation with smaller sizes (e.g., 64). Larger batches stabilized gradients and improved training efficiency.
   - Epochs: Training was capped at 20 epochs, with Early Stopping implemented to halt training if validation loss did not improve for 3 consecutive epochs.
   - Regularization:
     - Dropout: Added with a 30% probability to prevent overfitting.
     - L2 Regularization: Applied to dense and LSTM layers to add a penalty for overly complex models.
2. Model Architecture:
   - Encoder: The pre-trained VGG16 model was used for feature extraction, leveraging its proven performance on image datasets.
   - Decoder: The LSTM-based decoder was designed to handle sequential text generation, with an embedding size of 64 and 256 LSTM units.
3. Callbacks:
   - ReduceLROnPlateau: Dynamically reduced the learning rate when validation loss plateaued, ensuring better convergence.
   - EarlyStopping: Prevented overfitting by halting training once validation loss ceased to improve.

4.3 Discussion

1. Challenges and Observations:
   - Low BLEU Scores: The average BLEU score was low, likely due to the model's difficulty in capturing complex linguistic structures or diverse image features.
   - Validation Loss Plateau: The plateau in validation loss suggests either model capacity limitations or the need for additional data diversity.
   - Limited Dataset: The relatively small size of the Flickr8k dataset may have restricted the model's ability to generalize effectively.
2. Comparison with Existing Work:
   - State-of-the-art models like "Show, Attend, and Tell" use attention mechanisms to dynamically focus on relevant parts of the image. Incorporating attention could significantly improve performance.

  - Transformer-based architectures (e.g., Vision Transformers or GPT) have achieved much higher BLEU scores, particularly on larger datasets like COCO.

3. Potential Improvements:

   - Attention Mechanisms: Adding attention would allow the model to focus on specific image regions, improving caption relevance and accuracy.
   - Dataset Augmentation: Increasing dataset size or introducing augmentation techniques could help the model learn a wider range of visual and linguistic features.
   - Model Architecture: Transitioning to Transformer-based architectures could dramatically improve the model's ability to handle complex captions.
   - Pre-training: Pre-training the decoder on a large text corpus could help capture linguistic nuances, reducing the burden on the image-text alignment process.

4. Speculation on Future Directions:

   - Integrating multimodal learning techniques to handle both text and image data simultaneously could enhance model robustness.
   - Exploring low-resource solutions or lightweight architectures for real-world deployment in applications like accessibility tools.

5 Caption Generator Interface

To enhance the usability of the project, a Streamlit-based user interface was developed for generating captions from user-uploaded images.

Features:

- Model Integration: The trained captioning model and pre-trained VGG16 feature extractor are seamlessly integrated.

- Google Drive Support: Necessary model weights and tokenizer files are downloaded dynamically from Google Drive.

- User-Friendly Interface:

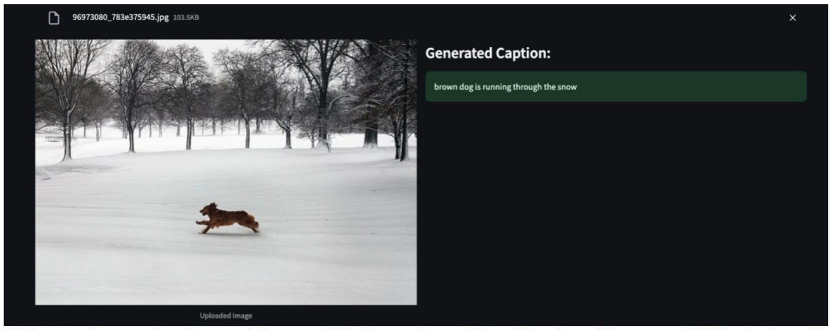
  - Users upload an image, and the app displays the uploaded image along with the generated caption.
  - Simple and intuitive layout for non-technical users.

Streamlit Code:

- Feature Extraction: The uploaded image is resized, preprocessed, and passed through VGG16 to extract visual features.

- Caption Generation: The LSTM-based decoder generates captions, token-by-token, until the <endseq> token is reached.

- Interactive UI: A sidebar allows users to select options and upload images dynamically.



Generated Caption:

brown dog is running through the snow

Figure 3: This figure showcases the interface of an image captioning model indicating its ability to generate a caption for a given image. In this example, the model accurately captions the image as "brown dog is running through the snow," demonstrating its capability to accurately analyze visual content and produce meaningful textual descriptions.

6 Conclusion

This project successfully demonstrated the application of an Encoder-Decoder framework for image caption generation. Using VGG16 for feature extraction and an LSTM-based decoder, the model was able to generate contextually relevant captions for a range of images. Additionally, we developed a model using the InceptionV3 pre-trained architecture, but it did not perform as well on the given dataset when compared to the VGG16 model. While the BLEU scores highlight areas for improvement, the results are promising and provide a solid foundation for future work.

With improvements in the attention mechanisms,improved datasets, and using advanced architectures, the model holds the potential to achieve state-of-the-art performance. Furthermore, this research opens up possibilities for impactful applications, such as assisting visually impaired individuals, by building a voice-enabled systems that reads out the generated captions, significantly enhancing the accessibility and making their lives easier. Hence, This work underscores the importance of combining computer vision and natural language processing techniques to solve complex multimodal tasks, paving the path for further innovations in this discipline.

References

[1]Shan-e-Fatima,Kratika Gupta, Deepti Goyal,Suman kumar Mishra(2024), Image Caption Generation using Deep learning Algorithm, DOI:10.53555/kuey.v30i5.4311.

[2]Verma, A., Yadav, A. K., Kumar, M., Yadav, D. (2024). Automatic image caption generation using deep learning. Multimedia Tools and Applications, 83(2), 5309-5325.

[3]Brooks, T., Holynski, A., Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18392-18402).

[4]Jabeen, S., Li, X., Amin, M. S., Bourahla, O., Li, S., Jabbar, A. (2023). A review on methods and applications in multimodal deep learning. ACM Transactions on Multimedia Computing, Communications and Applications, 19(2s), 1-41.

[5]Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D. (2024). Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems, 36.

[6]Satti Satish Kumar,Goluguri Nv, Prasad Maddula,Ravipati, N(2023),Image Caption Generation using ResNET-50 and LSTM, DOI - 10.1109/SILCON59133.2023.10404600.

[7]C. S. Kanimozhiselvi, K. V, K. S. P and K. S, "Image CaptioningUsing Deep Learning," 2022 International Conference on ComputerCommunication and Informatics (ICCCI), Coimbatore, India, 2022, pp.1-7, doi: 10.1109/ICCCI54379.2022.9740788.

[8]J. Sudhakar, V. V. Iyer and S. T. Sharmila, "Image Caption Generationusing Deep Neural Networks," 2022 International Conference for Ad-vancement in Technology (ICONAT), Goa, India, 2022, pp. 1-3, doi:10.1109/ICONAT53423.2022.9726074.

[9]P. Voditel, A. Gurjar, A. Pandey, A. Jain, N. Sharma and N. Dubey, "Image Captioning - A Deep Learning Approach Using CNN and LSTM Network," 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 343-348, doi: 10.1109/ICPCSN58827.2023.00062. keywords: Measurement;Deep learning;Training;Computer vision;Image recognition;Computational modeling;Computer architecture;CNN;LSTM;BLEU;VGG16;Image captioning;Deep learning

[10]G. Sairam, M. Mandha, P. Prashanth and P. Swetha, "Image Captioning using CNN and LSTM", 4th Smart Cities Symposium (SCS 2021), pp. 274-277, 2021.