Report On

Image Caption Generation with Enhanced Visual Attention Mechanisms (Adaptive Attention)

By Rahul Vinay

ABSTRACT

This project explores the use of adaptive attention mechanisms for image captioning, a task that bridges vision and language by generating context-aware textual descriptions of images. Unlike traditional RNN-based models with static attention, the proposed method employs an adaptive approach to dynamically decide whether to focus on image regions or rely on language context at each time step. The model integrates a ResNet50-based encoder, a sentinel gate-based attention module, and an LSTM decoder. The Flickr8k dataset is used for training and evaluation, with BLEU scores assessing performance. Despite achieving a baseline BLEU score of ~18.5, the work highlights significant learning opportunities and potential for future improvements. Challenges like dataset size limitations and BLEU's insensitivity to context have been addressed, with plans for extensions to larger datasets, alternative metrics, and spatial encodings. This project represents a foundational step in creating more efficient and accurate image captioning systems.

INTRODUCTION

Problem Statement:

Generating accurate captions for images requires a model to focus on the most relevant visual details while understanding the relationships between these details and the broader context. This task is challenging due to the intricate interplay of visual elements and the sequential nature of language generation. Static attention mechanisms, commonly used in traditional models, fail to adapt dynamically to these complexities, limiting their effectiveness.

Importance of the problem:

Image captioning has widespread applications, including:

1. **Accessibility:** Helping visually impaired users understand visual content through textual descriptions.
2. **Enhanced Image Search:** Allowing advanced search capabilities by associating meaningful captions with images.
3. **AI Systems:** Improving machine understanding of visual data, which can be applied to content moderation, e-commerce, and social media platforms.

Developing a robust image captioning system with adaptive attention mechanisms can significantly enhance caption relevance, making AI systems more useful and interpretable.

Objectives:

This project aims to design an image captioning model that integrates an **adaptive attention mechanism** capable of dynamically deciding when to focus on specific image regions or rely on language context. The primary goals include:

- Implementing a ResNet50-based encoder for feature extraction.
- Enhancing attention with a sentinel gate to prioritize relevant information.

- Evaluating the model using the Flickr8k dataset and BLEU scores.

Challenges:

1. **Complexity of Attention**: Ensuring high-quality and contextually accurate attention without significant computational overhead.
2. **Integration of Adaptive Attention**: Seamlessly deciding when to rely on visual data versus language modeling.
3. **Quality vs. Efficiency**: Balancing improved caption quality with processing speed and scalability to larger datasets.

LITERATURE REVIEW

Existing Solutions

Traditional image captioning models often rely on Recurrent Neural Networks (RNNs) combined with static attention mechanisms to generate captions. Key approaches include:

1. **Soft Attention**: Distributes attention uniformly across the entire image, often missing nuanced relationships.
2. **Hard Attention**: Focuses on discrete image regions but is computationally expensive and difficult to train.
3. **Static Attention**: Applies fixed attention weights, which fail to adapt dynamically to complex image contexts.

Limitations of Current Approaches

While these methods have achieved moderate success, they suffer from significant shortcomings:

- **Lack of Adaptability**: Static and soft attention mechanisms cannot dynamically adjust focus based on caption progression.
- **Computational Inefficiency**: Hard attention increases the processing burden, making it unsuitable for large-scale or real-time applications.
- **Loss of Contextual Nuance**: Uniform or rigid attention methods often overlook subtle yet critical relationships within the image.

Advancements and Contributions:

This project builds upon the seminal work *Show, Attend, and Tell* by Xu et al., which introduced attention mechanisms for image captioning.

The proposed model extends this work by:

1. Incorporating **adaptive attention** with a sentinel gate, enabling dynamic focus adjustments.
2. Using a **ResNet50 encoder** to extract high-dimensional feature maps for robust image representation.
3. **Preprocessing and Tokenization**: Built a vocabulary using frequency thresholds to exclude rare words, standardized captions by lowercasing and removing special characters, and introduced reserved tokens (<SOS>, <EOS>, <PAD>, <UNK>) to ensure consistency in caption representation.
4. **Training Pipeline Optimization**: Used advanced techniques such as sorting captions by length for efficient padding, leveraging GPU acceleration with batch processing, and optimizing model training with the Adam optimizer for improved convergence.
5. Emphasizing future scalability to larger datasets (e.g., Flickr30k, MS COCO) and exploring extensions like positional encodings for spatial awareness.

METHODOLOGY

Framework Overview:

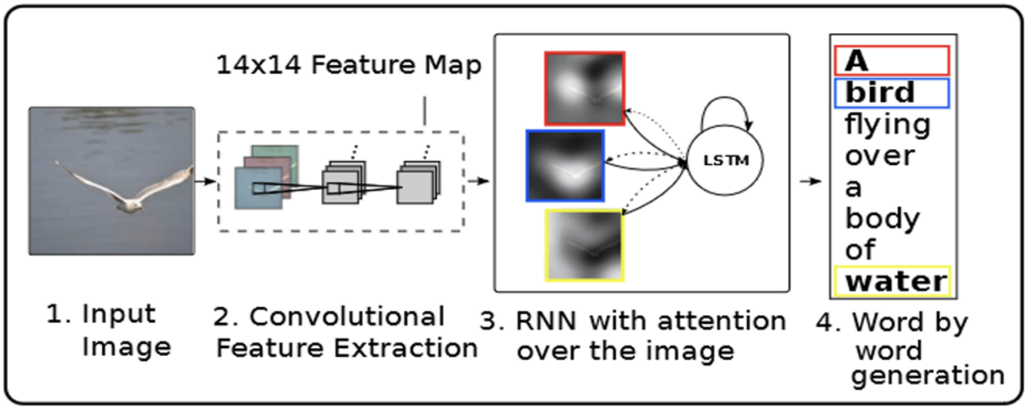

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figure 1: Image caption generation with attention. The model extracts a 14x14 feature map, applies attention through an RNN to focus on relevant regions, and generates a caption word by word.

The proposed model consists of three primary components:

1. **Encoder:** A ResNet50-based Convolutional Neural Network (CNN) extracts a 14×14 feature map from each input image. Each feature represents a distinct region of the image.

2. **Adaptive Attention Mechanism:** This module integrates a sentinel gate that dynamically decides when to focus on image regions or rely on prior language context, ensuring contextually relevant caption generation.

3. **Decoder:** An LSTM-based Recurrent Neural Network (RNN) generates captions word by word, conditioned on the attention-enhanced context vectors and previously generated words.

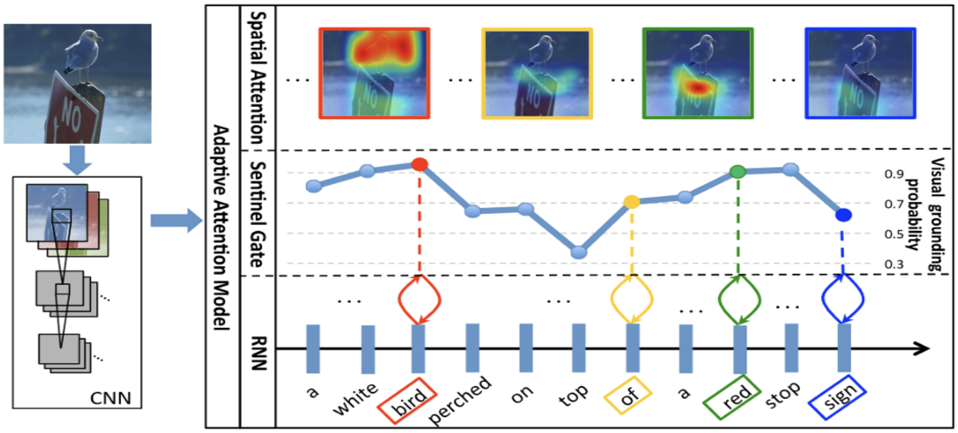


Figure 2: Sample adaptive attention model with a sentinel gate, dynamically deciding whether to attend to image regions or rely on context for each word in the caption.

Data Preprocessing and Tokenization:

- The Flickr8k dataset was used, consisting of 8,000 images and five captions per image. The dataset was downloaded via Kaggle API and organized into Images and captions.txt.

  Dataset can be downloaded manually from here if required: https://www.kaggle.com/datasets/adityajn105/flickr8k)

- Captions were tokenized using the NLTK library, and a vocabulary was built by applying a frequency threshold to exclude rare words.

- Reserved tokens (<SOS>, <EOS>, <PAD>, <UNK>) were introduced for structured processing, while captions were standardized (lowercasing, removing special characters).

Training Pipeline Optimization:

- Captions were sorted by length within each batch to reduce unnecessary padding.

- Batch processing was employed to efficiently leverage GPU acceleration during training.

- The Adam optimizer with a learning rate of 0.001 was used for parameter updates, ensuring stable convergence.

Evaluation Metrics

- **BLEU Scores**: Captions were evaluated using BLEU scores to assess n-gram overlap between generated and reference captions. The achieved BLEU score of ~18.5 serves as a baseline for future enhancements.

IMPLEMENTATION

Dataset Transformations

Preprocessing:

- Captions are tokenized using NLTK, standardized by converting to lowercase, and cleaned by removing special characters.
- A vocabulary is built using a frequency threshold of 3, excluding rare words. Reserved tokens (<SOS>, <EOS>, <PAD>, <UNK>) are included for structured sequence handling.
- Images are resized to 256×256, randomly cropped to 224×224, and normalized using ImageNet statistics (mean: (0.485, 0.456, 0.406), standard deviation: (0.229, 0.224, 0.225)).
- Random horizontal flipping is applied with a probability of 0.5 for data augmentation.

Dataset Representation:

- Captions are numericalized by replacing tokens with their corresponding indices from the vocabulary. Sequences are padded to the maximum length in the batch to ensure uniform input dimensions.

DataLoader:

- A custom collation function is implemented to sort captions by length within each batch, minimizing padding overhead and optimizing GPU utilization during training.

Model Architecture

Encoder (CNN):

- A pretrained ResNet50 model is employed, with the final fully connected layers removed to retain spatial feature maps.

- Adaptive average pooling ensures consistent feature map dimensions of [batch_size, num_pixels, encoder_dim], where encoder_dim=2048 and num_pixels=14×14=196.

Attention Mechanism:

- Dynamically computes attention weights using linear transformations applied to the encoder's output features and the decoder's hidden state.
- Produces a weighted combination of the encoder's features, emphasizing the most relevant regions for each word in the caption.

Decoder (LSTM):

- An embedding layer maps input tokens to vectors of size embed_dim=256.
- The LSTM cell generates captions word-by-word, initialized with hidden and cell states computed from the encoder's output.
- A sentinel gate dynamically prioritizes contextually impor tant features, enhancing the relevance of generated captions.

Output Layer:

- The fully connected output layer generates vocabulary probabilities for the next word in the sequence, with predictions guided by a softmax activation.

Training Pipeline

Loss Function: Cross-entropy loss, ignoring <PAD> tokens in the target captions, is used to optimize the model.

Optimizer: Adam optimizer is employed with a learning rate of 0.001, ensuring stable and efficient parameter updates.

Batch Processing: Captions are sorted by length to reduce padding, and data is processed in batches to leverage GPU acceleration.

Evaluation Workflow

- BLEU scores are computed to measure n-gram overlap between the generated and reference captions, with results serving as a baseline.
- Sample outputs are visualized by pairing input images with their generated captions to provide qualitative insights into model performance.

RESULTS

Quantitative Analysis

BLEU Score:

- The model achieved a BLEU score of ~18.5 on the Flickr8k dataset, which serves as a baseline for future improvements.
- BLEU was used to evaluate the n-gram overlap between generated captions and human-annotated references.

Limitations of BLEU:

- Insensitive to synonyms or contextual meaning.
- Overly favors shorter captions, potentially undervaluing longer, well-formed sentences.

Qualitative Analysis

Generated Captions:

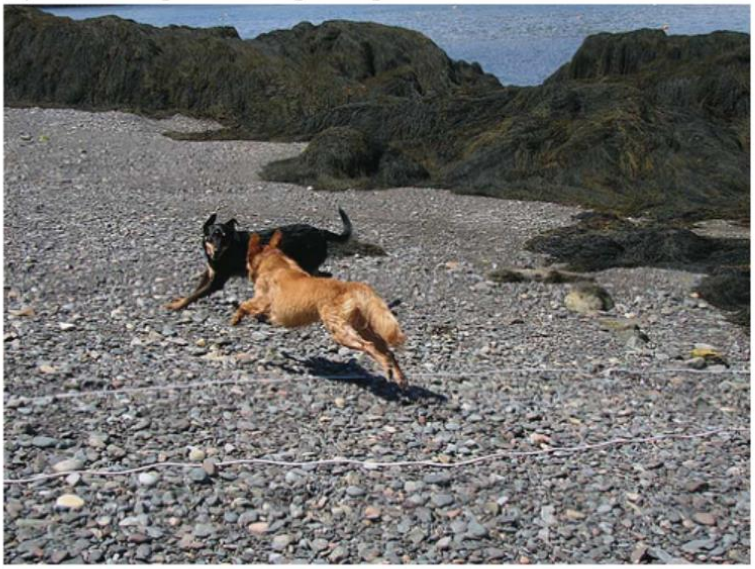
The model demonstrated the ability to generate contextually relevant captions for simpler images, such as:

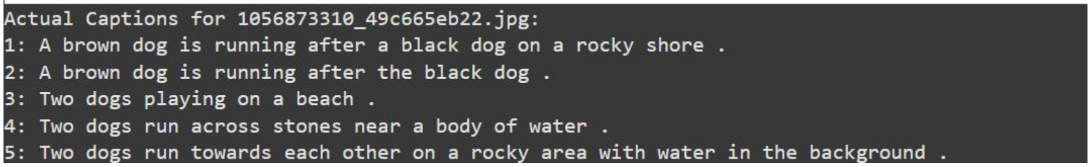
Input: An image of a dog playing in the garden.

Generated Caption: "A dog is playing in the garden."

For complex scenes, the model occasionally produced incomplete or generic descriptions, highlighting areas for improvement.

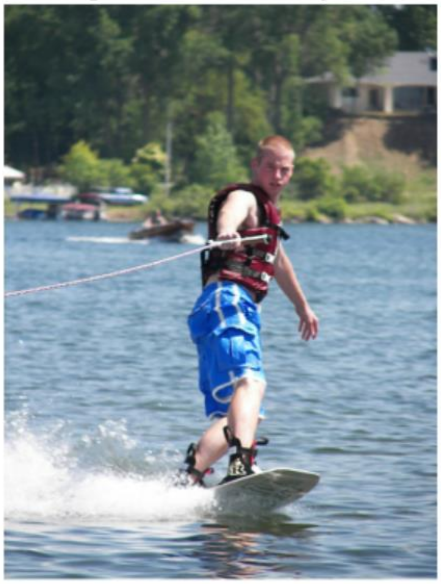
Generated Caption: a dog is running along a beach with a stick in its mouth <EOS>

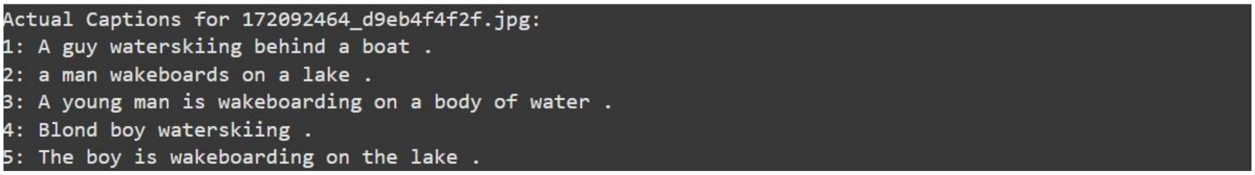




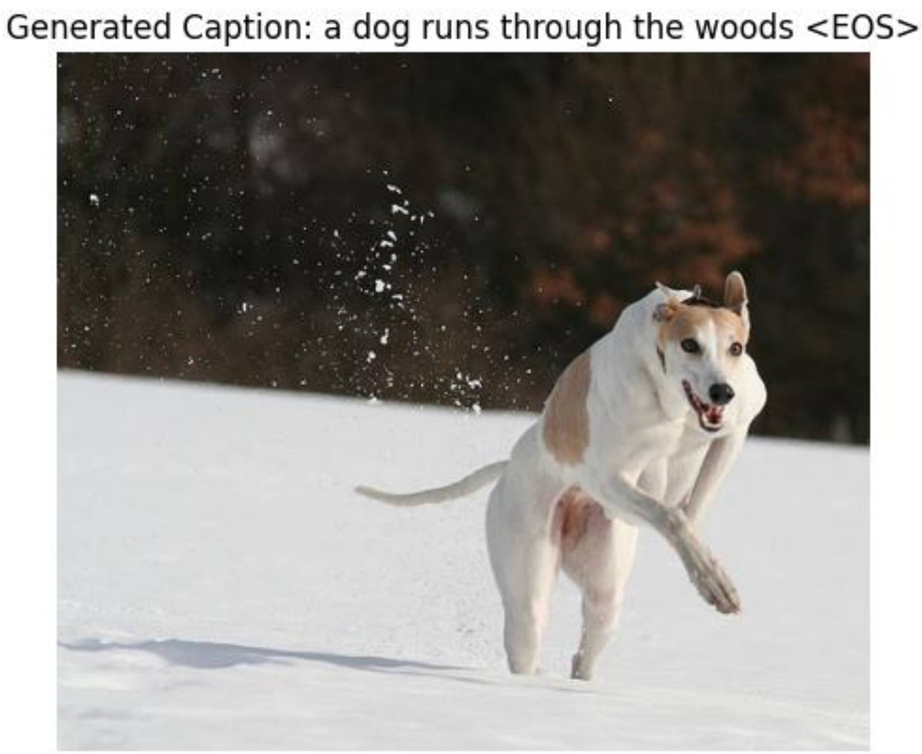
```
Actual Captions for 1056873310_49c665eb22.jpg:
1: A brown dog is running after a black dog on a rocky shore .
2: A brown dog is running after the black dog .
3: Two dogs playing on a beach .
4: Two dogs run across stones near a body of water .
5: Two dogs run towards each other on a rocky area with water in the background .
```

Generated Caption: a man is standing in the ocean fully clothed holding a fishing pole <EOS>

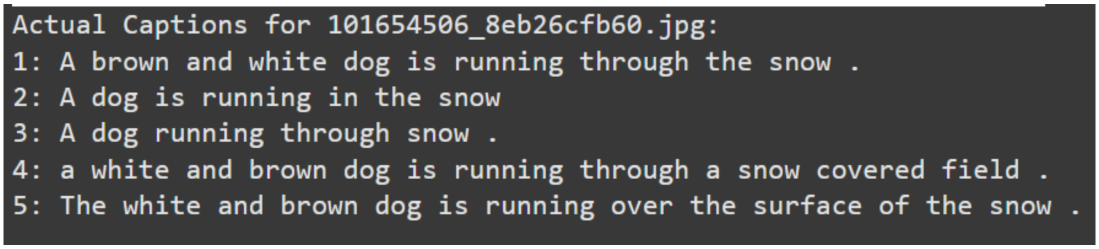




```
Actual Captions for 172092464_d9eb4f4f2f.jpg:
1: A guy waterskiing behind a boat .
2: a man wakeboards on a lake .
3: A young man is wakeboarding on a body of water .
4: Blond boy waterskiing .
5: The boy is wakeboarding on the lake .
```



Generated Caption: a dog runs through the woods <EOS>



```
Actual Captions for 101654506_8eb26cfb60.jpg:
1: A brown and white dog is running through the snow .
2: A dog is running in the snow
3: A dog running through snow .
4: a white and brown dog is running through a snow covered field .
5: The white and brown dog is running over the surface of the snow .
```

Observations

Strengths:

- Demonstrated the ability to generate grammatically correct captions.
- Adaptive attention improved focus on relevant image regions during simpler scenes.

Limitations:

- Performance is constrained by the size of the Flickr8k dataset.
- Challenges in generating diverse and detailed captions for complex or crowded scenes.
- BLEU's shortcomings in evaluating semantic meaning limit its effectiveness as the sole metric.

DISCUSSION

Learning Outcomes:

- The project validated the effectiveness of adaptive attention mechanisms.
- Sentinel gates introduced flexibility in deciding when to prioritize image context over language modeling.
- Practical experience was gained in integrating CNNs and RNNs, managing dataset preprocessing, and implementing evaluation workflows.

Future Potential:

- Extend the model to larger datasets (e.g., Flickr30k, MS COCO) to improve performance.
- Explore alternative evaluation metrics, such as METEOR or CIDEr, to capture contextual and semantic nuances.
- Incorporate positional encodings or object-detection features for enhanced spatial understanding.

REFERENCES

[1]. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Proceedings of the 32nd International Conference on Machine Learning, 2015.

[2]. Kaggle Flickr8k Dataset: https://www.kaggle.com/datasets/adityajn105/flickr8k.

[3] Lu, J., Xiong, C., Parikh, D., & Socher, R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. Salesforce Research, Virginia Tech, and Georgia Institute of Technology.

[4] Rister, B., Lawson, D., & Al-Malla, M. A. Image Captioning with Attention. Stanford University.

[5] Al-Malla, M.A., Jafar, A., & Ghneim, N. Image Captioning Model Using Attention and Object Features to Mimic Human Image Understanding. Journal of Big Data, 9:20, 2022.

[6] Attention Mechanism for Caption Generation. Analytics Vidhya Blog, 2020. Available from https://www.analyticsvidhya.com/blog/2020/11/attention-mechanism-for-caption-generation/