Generating Image Captions Using Dynamic Mode Decomposition and CNN-LSTM Architecture

SANJANA MCS
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India
sanjana.machavolu@gmail.com

SIDDHARTH D
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India
dsiddharth25@gmail.com

ABINAYA N
School of Artificial Intelligence
Amrita Vishwa Vidyapeetham
Coimbatore, India
abinayanithyan2403@gmail.com

Dr.Sachin Kumar.S
Amrita School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
s_sachinkumar@cb.amrita.edu

***Abstract*—Image recognition is a complex task that relies on the intersection of computer vision and natural language processing. It requires a deep understanding of the visual content of images and the ability to create coherent and descriptive textual captions. In this work, we explore the application of Dynamic Mode Decomposition (DMD) for feature extraction in image captioning combined with a CNN-LSTM architecture. Despite the theoretical advantages of DMD in capturing dominant spatial and temporal features, our experiments demonstrate that this approach did not yield the desired improvements in caption generation. Specifically, the generated captions lacked accuracy and coherence, highlighting the limitations of DMD in this context. This paper provides an in-depth analysis of these challenges and suggests directions for future research.**

I. Introduction

Traditional approaches to image captioning primarily rely on convolutional neural networks (CNNs) to extract image features, which are then processed by recurrent neural networks (RNNs) to generate textual descriptions. Although these methods have made significant progress, they often overlook the potential benefits of decomposing image features into distinct components that capture different aspects of image structure. In this study, we first implemented a classical image captioning model using a CNN-LSTM architecture, which involved extracting visual features from images using a CNN and then feeding these features into an LSTM to generate captions.

In our proposed approach, we aimed to leverage the power of various color space representations, including YUV, YCbCr, and CIELab, to capture diverse visual features essential for accurate image captioning. We adapted DMD, renowned for its analytical prowess in uncovering underlying structures within dynamic systems, to decompose image data into low-rank and sparse components. This decomposition was intended to facilitate a deeper understanding of the image dynamics, enabling the extraction of salient features crucial for caption generation.

Despite these theoretical advantages, our practical implementation did not achieve the expected results. The decomposed components, processed separately through deep CNN models, did not contribute to the anticipated improvement in feature representation. The generated captions were often inaccurate and failed to capture the essence of the images, as illustrated by the example provided (see Figure 1).

The primary contributions of this paper are as follows:

- Critical Evaluation of DMD for Image Decomposition: We detail the application of DMD in the context of image captioning and analyze why it fell short in improving caption generation.
- Lessons Learned and Future Directions: We provide insights into the challenges faced and suggest potential avenues for further research to enhance image captioning models.

II. Literature review

In recent years, significant progress has been made in the field of image caption generation using computer vision and natural language processing (NLP) techniques. The integration of convolutional neural networks (CNN) and long short-term memory (LSTM) networks has been the cornerstone of many approaches in this area. For example, Afeefa Nazneen N Z and Dr. Shreedhara K S [1] proposed a model that combines CNN for extracting visual features and LSTM for generating textual descriptions. Their model is designed to ensure that the captions generated are relevant and limited to a pre-defined vocabulary, thereby increasing the accuracy and usability of the captions. However, the requirement for a significant amount of training data presents a significant challenge in terms of data collection and processing.

Similarly, another study by Dr. G. Lakshmi Vara Prasad and colleagues developed a CNN-LSTM model [2]to automatically recognize and describe images in English. Their approach effectively combines computer vision and NLP techniques to produce robust image labels. While the methodology benefits

from the use of pre-trained networks and large datasets, the study design lacks explicit details, potentially limiting its generalizability.

By extending these basic models, Songtao Ding and coworkers presented a new image caption model [3]based on high-level image features and a bottom-up attention mechanism. This model aims to mimic human visual attention mechanisms by efficiently combining low-level and high-level features to focus on relevant parts of an image. Despite its strengths in solving caption generation problems and achieving good performance on reference datasets, the method is somewhat limited by the limited variety of its descriptors and the difficulty of accurately describing images with complex backgrounds.

In their comprehensive review, Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar [4] discussed various deep learning methods for image description, including attention-based methods, Transformers, scene graphs, and visual language pretraining. Their review highlights the ability of these methods to capture complex relationships and produce natural language descriptions. However, issues such as exact word-image matching and bias in the training data have been noted, underscoring the need for better evaluation tools and more diverse datasets.

Another notable study by Smriti Sehgal, Jyoti Sharma, and Natasha Chaudhary [5]combined CNNs and RNNs with NLP techniques to develop an image caption generation model. This model aims to assist visually impaired individuals by automating caption generation. While the model effectively uses feature maps to capture image positions and efficiently organizes files, small variations in the input feature maps can affect consistency.

Ali Farhadi and colleagues investigated the potential of automatic methods of generating descriptive sentences from images [6]. Their approach introduced a new representational mediator between images and sentences and used a discriminative method for sentence annotation. Despite the complex nature of their model, the lack of a data set with matching sentences to evaluate and the difficulty of quantitative evaluation presented challenges.

Dynamic mode decomposition (DMD) has also been used in saliency detection, as shown by O. K. Sikha and coworkers [7]. Their method identifies salient regions in images by decomposing image data into lower-quality and sparse components, combining color and brightness information. Although the method exhibits competitive performance and computational efficiency, it is primarily focused on natural scenes and has an unoptimized MATLAB implementation.

In another hybrid approach, Aishwarya Maroju, Sneha Sri Doma and Lahari Chandarlapati used ResNet for image feature extraction and LSTM for caption generation [8]. Their model solves the vanishing gradient problem commonly encountered in traditional CNN-RNN models. However, the model requires considerable training data and iterations to effectively learn the complex relationships within the data.

Overall, the literature on image caption generation reveals a number of innovative approaches, each contributing to the field in a unique way. These studies highlight the potential and challenges of integrating computer vision and NLP techniques to automatically generate descriptive captions for images. Future research should focus on overcoming data limitations, improving the generalizability of the model, and increasing the accuracy and diversity of generated captions.

III. METHODOLOGY

*1) **Data Acquisition**:* For this study, the Flickr8k dataset is used for data acquisition. The Flickr8k dataset is a widely recognized benchmark in the field of image captioning and provides a substantial amount of data necessary for training and evaluating models. Key features of the Flickr8k dataset include:

- **Image Collection:** The dataset contains 8,000 images that have been carefully selected from the Flickr photo-sharing website. These images cover a diverse range of scenes, objects, and activities.
- **Captions:** Each image in the dataset is accompanied by five different captions. These captions were created by human annotators and provide a variety of descriptive phrases for the same image, capturing different aspects and perspectives.
- **Diversity:** The dataset includes images with a wide variety of subjects, including people, animals, objects, and natural scenes, making it well-suited for training models to generate captions for a broad range of image types.
- **Annotations:** The captions provided are detailed and cover multiple attributes of the images, such as actions, objects, and contextual information, which help in training models to generate comprehensive and contextually relevant descriptions.

*2) **Data Preprocessing**:*

- **Image Preprocessing:** Initially images are loaded from the dataset directory and resized to 256x256 pixels, which is the standard size for image preprocessing. After that, the images are transformed from RGB to YUV and YCbCr colour spaces in order to capture various visual characteristics. A comprehensive representation is produced by concatenating the YUV and YCbCr images along the final dimension.
- **Captions Preprocessing:** Text preprocessing involves loading and cleaning captions associated with each image, converting them to lowercase, and removing punctuation. A tokenizer is used to convert words into numerical representations, creating a vocabulary of the most frequent words. Captions are then encoded using techniques like n-grams, TF-IDF, and Bag of Words to capture context, word importance, and frequency. The tokenized captions are padded to ensure uniform length for neural network processing. Finally, the data is split into training and testing sets to allow for model training and performance evaluation on separate data portions. This comprehensive

preprocessing ensures the textual data is optimized for training the image captioning model.

*3) **Feature Extraction**:* Feature extraction initially involved using Dynamic Mode Decomposition (DMD) on the concatenated YUV and YCbCr images. DMD decomposes each image into low-rank and sparse components, effectively separating the essential structural information from the detailed nuances. However, this method did not enhance the feature extraction process as expected. The decomposed components were then processed through a pre-trained VGG16 model to extract high-level features. Despite the sophisticated decomposition, the DMD approach did not yield significant improvements in generating accurate captions, prompting the need for an alternative feature extraction strategy.

*4) **Model Training**:* The extracted features and preprocessed captions were used to train a CNN-LSTM model. The combined image features were intended to serve as inputs to the CNN, which captures spatial hierarchies in the data, while the LSTM handles the sequential nature of the text data. Specifically, the CNN processes the image features, and the LSTM generates the corresponding captions. The model was trained on the combined dataset of image features and encoded captions, optimizing for categorical cross-entropy loss using the Adam optimizer. However, the inclusion of DMD-derived features did not improve the model's performance, indicating the need for a more effective feature extraction method.

*5) **Evaluation**:* The BLEU score is used to evaluate the performance of the trained model. The BLEU score assesses how well the trained model performs in natural language processing by comparing generated text with reference captions. The BLEU scores can vary between 0 and 1, with higher scores signaling improved performance. The evaluation involves generating descriptions for a set of test images and assessing them against the correct captions using the BLEU metric. The results revealed that the DMD approach did not significantly enhance the BLEU scores, reflecting the model's limited capability in generating accurate and contextually appropriate image descriptions.

*6) **Explainability and Interpretability**:* To enhance the transparency of our image captioning model, we utilized techniques like Grad-CAM to identify the emphasized areas of the image during caption generation. By concentrating on these areas, we aimed to understand the model's decision-making process better and confirm that the descriptions are both appropriate and accurate. However, the explanations provided by Grad-CAM indicated that the features extracted via DMD did not align well with the critical regions of the images, further highlighting the limitations of using DMD in this context.

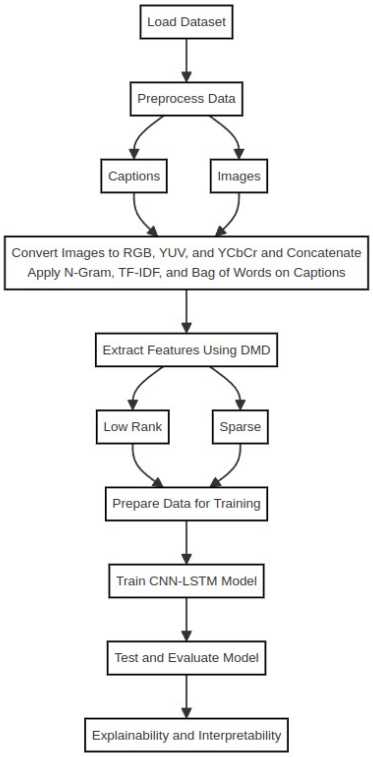
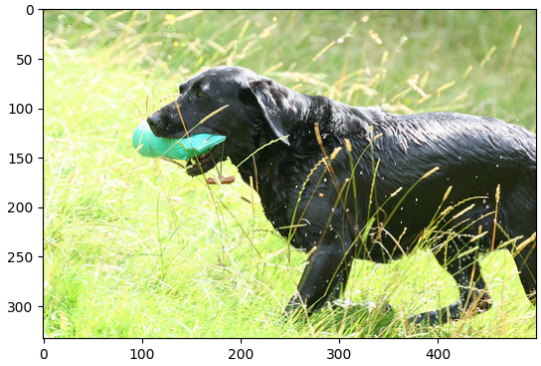


Fig. 1. BLOCK SCHEMATIC OF WORK FLOW

A. *Experimental Results*

1) Classic Approach Results: In our experiment, we initially used a basic image captioning model to generate captions for a set of images. The model was a simple neural network trained on a standard dataset of images and captions using the classic approach of a CNN-LSTM architecture. Here are the results for a few images:



- **Original Caption:** A black dog carries a green toy in his mouth as he walks through the grass.
- **Predicted Caption:** start black dog is running through the grass end

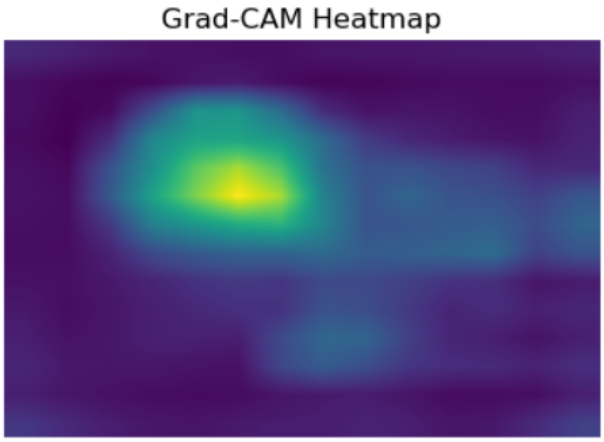
- **Original Caption:** A young boy rides on a surfboard with light blue water behind him.
- **Predicted Caption:** start man in wetsuit is wakeboarding on the water end

These results were obtained using the classic image captioning approach of employing a CNN-LSTM model. The predicted captions, while somewhat relevant, often lacked accuracy and detail, indicating the limitations of this method. This motivated us to explore the potential of Dynamic Mode Decomposition (DMD) for improving feature extraction and caption generation, though as noted earlier, this approach did not yield the desired improvements.

2) DMD Approach Results: We applied the Dynamic Mode Decomposition (DMD) technique for feature extraction, expecting it to enhance the accuracy of the generated captions. However, the results demonstrated that DMD did not contribute to improved caption quality.

- **Original Caption:** Black dog is running through the grass.
- **Predicted Caption:** a a

These results indicate that while DMD has potential in other domains, it did not prove effective for the task of image captioning in this experiment. Alternative methods for feature extraction and caption generation need to be explored to achieve better results.



B. Discussions

In this study, we explored an image captioning approach by combining multiple color space representations (RGB, YUV, and YCbCr) with Dynamic Mode Decomposition (DMD) for feature extraction. Our aim was to improve the accuracy and contextual relevance of generated captions by leveraging these complex features within a CNN-LSTM model. However, contrary to our expectations, the results indicated that DMD did not enhance the caption quality and, in fact, introduced several limitations.

One of the primary reasons for the failure of DMD in generating accurate captions is that it decomposes images into low-rank and sparse components. While this decomposition can highlight important structures, it also leads to a significant loss of contextual and detailed information. This lack of comprehensive image features is detrimental to the caption generation process, which relies heavily on capturing a wide range of details to produce meaningful and coherent descriptions.

Additionally, the computational complexity associated with DMD and the subsequent feature extraction poses challenges for practical applications, particularly in real-time scenarios. These findings highlight the limitations of using DMD for image captioning and underscore the need for alternative methods that can better preserve and utilize the full spectrum of image information.

C. Conclusions

This paper examined the integration of Dynamic Mode Decomposition (DMD) with a CNN-LSTM framework for image captioning. Our approach aimed to utilize DMD's ability to decompose images into low-rank and sparse components to capture both primary features and intricate details. However, the results revealed that this decomposition process leads to a significant loss of vital contextual information, which is crucial for generating accurate and detailed captions.

The experimental results showed that the use of DMD did not improve caption generation. Instead, it resulted in incomplete and often incorrect captions. This outcome underscores the limitations of DMD in feature extraction for image captioning. The classic CNN-LSTM model, without DMD, performed better by maintaining a more holistic representation of the images.

The findings from our experiments on the Flickr8k dataset indicate that DMD is not suitable for generating image captions due to its inherent limitation of losing detailed contextual information. These results suggest that alternative methods that preserve the comprehensive features of images are necessary for improving the accuracy and richness of generated captions.

D. Justifications

- The primary justifications for the claim that DMD is not suitable for image captioning are as follows:
- **Loss of Contextual Information:** DMD focuses on decomposing images into low-rank and sparse components, which strips away essential contextual details necessary for generating meaningful captions.
- **Insufficient Feature Representation:** The features extracted by DMD do not capture the full complexity and richness of the image content, leading to incomplete and inaccurate captions.
- **Incompatibility with Sequential Models:** The sequential nature of LSTM models requires comprehensive and detailed features to generate coherent text. DMD's decomposition process fails to provide such features, resulting in poor performance.
- **Computational Complexity:** The process of applying DMD and subsequent feature extraction is computationally intensive, making it impractical for real-time applications and not justifying the potential benefits.
  These points collectively demonstrate that DMD is not an effective technique for feature extraction in the context of image captioning.

E. Future Scope

Future research should focus on validating our findings with larger and more diverse datasets, such as Flickr30k or MS COCO, and explore alternative methods for feature extraction that preserve the full range of image information. Advanced attention mechanisms or transformer-based models could be integrated to further improve performance. Additionally, optimizing the computational efficiency of the feature extraction process will be essential for practical real-time applications. This exploration could lead to the development of more effective and efficient methods for automatic image caption generation, overcoming the limitations identified with DMD.

REFERENCES

[1] Nazneen N Z, Afeefa, & K S, Shreedhara. (2022). Image Caption Generation using Convolutional Neural Network and Long Short Term Memory. *International Transaction on Electrical Energy Systems*, 11(4). Retrieved from http://www.iteejournal.org/v11no4august22_pdf2.pdf

[2] Prasad, G. Lakshmi Vara, et al. (Year). Image Caption Generator Using CNN and LSTM. *Journal Name*, Volume(Issue). Retrieved from https://www.google.com/url?sa=i&url=https%3A%2F%2Fsajet.in%2Findex.php%2Fjournal%2Farticle%2Fdownload%2F219%2F226&psig=AOvVaw2jWc-feb04Pb4atRTNccJR&ust=1716889946521000&source=images&cd=vfe&opi=89978449&ved=0CAcQrpoMahcKEwjw1NzRx62GAxUAAAAAHQAAAAAQBA

[3] Ding, Songtao, et al. (Year). Image caption generation with high-level image features. *Journal Name*, Volume(Issue). Retrieved from https://www.sciencedirect.com/science/article/pii/S0167865519301047

[4] Ghandi, Taraneh, Pourreza, Hamidreza, & Mahyar, Hamidreza. (Year). DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW. *arXiv preprint arXiv:2201.12944*. Retrieved from https://arxiv.org/pdf/2201.12944

[5] Sehgal, Smriti, Sharma, Jyoti, & Chaudhary, Natasha. (Year). Generating Image Captions based on Deep Learning and Natural language Processing. *IEEE Xplore*. Retrieved from https://ieeexplore.ieee.org/document/9197977

[6] Farhadi, Ali, et al. (Year). Every Picture Tells a Story: Generating Sentences from Images. *Journal Name*, Volume(Issue). Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-15561-1_2

[7] Sikha, O. K., et al. (Year). Salient Region Detection and Segmentation in Images using Dynamic Mode Decomposition. *Journal Name*, Volume(Issue). Retrieved from https://www.researchgate.net/publication/305186354_Salient_Region_Detection_and_Segmentation_in_Images_using_Dynamic_Mode_Decomposition

[8] Maroju, Aishwarya, Doma, Sneha Sri, & Chandarlapati, Lahari. (Year). Image Caption Generation Using Deep Learning Technique. *IEEE Xplore*. Retrieved from https://ieeexplore.ieee.org/document/8697360