Multimodel Narrative Generation: Deep learning Integration with NLP for Visual Story Telling and Image Captioning

Deepthi Godavarthi[1], Javvaji Akshita[2], Ghorakavi Venkata Naga Lakshmi Saranya[3]

[1] VIT-AP University, Andhra Pradesh, India
[2] VIT-AP University, Andhra Pradesh, India
[3] VIT-AP University, Andhra Pradesh, India

deepthi.g@vitap.ac.in
akshita.21bce9040@vitapstudent.ac.in
saranya.21bce9084@vitapstudent.ac.in

Abstract:

Right now, there has been considerable research on application of artificial intelligence (AI) for generating illustrative captions for images and story generation based on caption generated. However, a common constraint generally onbserved in is no clarity and correctness of generated captions and the limited number of variations they provide. To solve this problem, our study introduces a sophisticated framework that include encoder-decoder architecture for the generation of short story by utilizing a manually classified corpus story dataset and a well-known image caption dataset.[12] There are three main functions in this manuscript. First, we provide an innovative unsupervised deep learning framework, which integrated recurrent neural networks (RNNs) with an encoder-decoder model. Second, we introduced a substantial story corpus, carefully put together and validated, comprising different genres[15]. As such, our exhibited framework not only holds stability in stimulating enhanced AI-based storytelling but also is upholder for the practical usage of our model in various real world applications aimed at helping human storytellers in discovering novel narrative concepts.[2]

Introduction:

Recent years have witnessed significant AI advancements, extending into creative realms like music composition and storytelling. Our study focuses on using AI to generate image captions and craft short stories, aligning with input image contexts for enhanced engagement. We tackle the "visual story writer" challenge with unsupervised sequence-to-sequence learning, utilizing a GRU architecture. Our framework incorporates predicted image captions to generate cohesive stories across genres. By mapping sentences to skip-thought vectors and employing a visual-semantic embedding model, we ensure narrative coherence. The study comprises six sections, covering introduction, caption generation trends, dataset description, model architecture, experimental evaluation, and future directions.

A. IMAGE CAPTION GENERATION

For Generating captions two approaches exist: retrieval-based and template-based methods. The former retrieves captions from similar images, while the latter completes predefined templates using identified objects and interactions. Recent advancements in deep learning have led to frameworks using benchmark datasets like COCO and Deja Image Captions, focusing on multimodal learning, encoder-decoder frameworks, compositional architectures, and attention-guided networks. Pretrained models like VGG or AlexNet extract image features, which are then processed by caption generators using NLP methods and RNNs.

B. DATASET DESCRIPTION

The study utilizes two datasets: (1) Books from Smashwords, containing stories with over 20,000 words to minimize noise, comprising 550 romance and 72 horror stories. (2) The Conceptual Captions dataset by Google, introduced in 2018

BOOK DATASET

To download books in PDF format we made use of an automated Python crawler which was followed by the usage of the python pdf2txt library to convert the already present text into plain text[11]. Subsequently, data preprocessing is done using a Python NLP library to remove excess and unnecessary blank rows, and all other contents which were put together into a single file.

CONCEPTUAL CAPTIONS DATASET

All sentences in the dataset were converted to lowercase, and non-alphanumeric characters were discarded. Additionally, the words which were appearing fewer than five times in the dataset were filtered out[13].

Literature Survey:

Recent research in multimedia content retrieval aims to improve retrieval effectiveness and indexing by considering narrative perception. For instance, [7] proposed a system that arranged crowdfunding plot synopses with images to enable story-style content retrieval, utilizing a function to compare sentences and incorporate personal identities and keywords generated from captions. In another study [8], authors integrated a neural network structure using pairwise and single element-based predictions, leveraging text and image features to capture captivating temporal features. Gaur [9] investigated hashtags from images to derive meaningful stories, employing a dual encoder-decoder model to generate hashtags and a multi-layer RNN for paragraph generation. Similarly, Alexander [2] developed a model capable of generating visually descriptive captions by separating style and semantics using NLP techniques, implementing frame semantics and merged language-based models for sentence generation.

Our work differs from prior research by focusing on generating accurate stories from image captions, which describe the main content in an image. Additionally, we introduced a three-step method for creating stories with style transfer based on skip-thought modelling [14].

In [13] This work contributed to understanding the repetitiveness in image descriptions and provided a dataset with expressive descriptions.

In [16] This paper proposed an encoder-decoder framework for image captioning, where a CNN is used to encode image features and an LSTM network is employed to generate captions. It introduced the concept of using neural networks for end-to-end caption generation.

In [14] The introduction of deep convolutional neural networks (CNNs), such as the VGG16 architecture, revolutionized image recognition tasks by enabling the extraction of high-level features from images.

In [12] This paper introduced the task of describing images using natural language, laying the foundation for image captioning research.

In [8] Addressing the issue of generating diverse captions, this paper proposed a discriminability loss function to encourage diversity while ensuring discriminative quality. The approach facilitated the generation of varied and semantically meaningful captions for images.

Methodology:

This section of the research paper elaborates on the proposed visual story writer model. The model comprises three core components: firstly, the training of both the visual-semantic alignment model and the skip-thought encoder on the conceptual captions dataset; secondly, the extraction of skip-thought vectors from the collected stories dataset; and thirdly, the execution of deep style transfer to morph the caption style into the story style utilizing the skip-thought decoder. Moving on to the image caption generation process, it hinges on the deep visual-semantic alignments model [7], which consists of an encoder and a decoder. Initially, a region convolutional neural network (RCNN) is employed to establish mappings between image regions and words, thereby creating a unified, multimodal embedding on the encoder side. Subsequently, a recurrent neural network (RNN) is applied on the decoder side to craft embedding representations, ensuring proximity between semantically linked concepts across both words and image regions, thereby enriching the narrative potential of the visual storytelling process.[5]

IMAGE CAPTON GENERATION:

A. IMAGE REPRESENTATION:

The RCNN model in this study underwent pre-training on the ImageNet dataset, consisting of

1000 object classes, and subsequent fine-tuning on a subset of 200 object classes from the ImageNet Object Localization Challenge. The RCNN model comprises approximately 58 million parameters and transforms the pixels within localized bounding boxes into a 2048-dimensional activation via a fully connected layer.

B. SENTENCE REPRESENTATION:

The RNN model processes a sequence of N words, converting them into an h-dimensional vector enriched by surrounding context. The Skip-Gram Sentence Encoder-Decoder employs a bi-directional skip-gram model to learn word representations based on distributional theory, encoding sentences into fixed vectors.

C. ENCODER:

In image captioning, Convolutional Neural Networks (CNNs) are used for feature extraction, while in text-centric tasks like story generation, Recurrent Neural Networks (RNNs) or Transformer models process sequential input data. The encoder extracts features and performs dimensionality reduction to create fixed-length representations preserving semantic information. The encoder's operation can be represented by the equation:

$$H_t = f(W_e.M_t+W_f.h_{t-1}+W_b.h_{t+1}+b)$$

Where; H_t represents the hidden state at time step

M_t is a column vector representing the t-th word in a word vocabulary,

W_e, W_f and W_b are weight matrices,

f is the activation function, and

b is the bias term.

D. DECODER:

The decoder generates sequential output based on encoded representations from the encoder, leveraging conditional generation and language modeling techniques. Attention mechanisms focus on relevant features, and during inference, beam search enhances output fluency. Through iterative refinement during training, the decoder learns to generate coherent and contextually relevant output sequences, essential for successful image captioning and story generation projects. The decoder's operation can be represented by the equation:

$$S_t = f(W_e.M_t+W_f.h_{t-1}+W_b.h_{t+1}+b)$$

Different models for image captions:

VGG-16:

Features: Renowned for high-level visual feature extraction through hierarchical layers of convolutional and max-pooling layers.

Transfer Learning: Utilizes pre-training on ImageNet, enabling adaptation of learned features to specific tasks like image captioning and story generation.

AlexNet:

Architecture: Known for its success in the ImageNet Challenge, comprising convolutional and max-pooling layers for feature learning.

Feature Learning: Extracts descriptive visual features essential for generating detailed captions or narratives.

Transfer Learning: Can be fine-tuned on task-specific data for enhanced performance in generating captions or stories.

GoogLeNet (Inception-v1):

Innovation: Introduces inception modul for capturing features at multiple scales efficiently.

Efficiency: Designed for computational efficiency while maintaining robust feature extraction capabilities.

Transfer Learning: Pre-trained on ImageNet, adaptable to specific datasets for improved performance in tasks like image captioning and storytelling.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312

Total params: 134260544 (512.16 MB)
Trainable params: 134260544 (512.16 MB)
Non-trainable params: 0 (0.00 Byte)

None

Figure 1: Vgg-16 model weights excluding the last fully connected layer

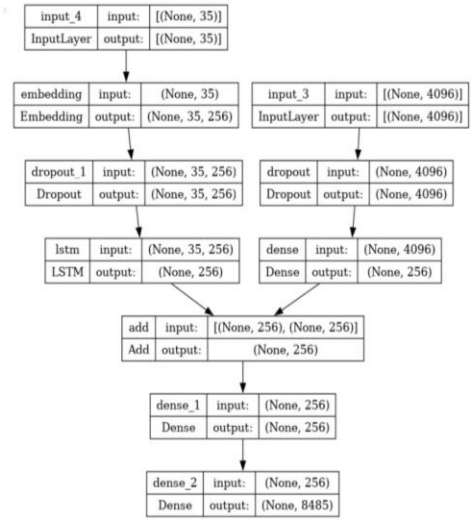

Figure 2 : clear flow chart of model depicting encoder layers, image feature layers and sequence feature layers and decoder layers.

Description of Algorithm for story telling:

Load the trained encoder and decoder models from their respective file paths. Import necessary libraries such as pandas, numpy, tensorflow, etc. Load captions from the dataset using pandas DataFrame. Tokenize the captions to create a vocabulary and map each word to an index. Save the tokenizer as a JSON file for future use. Define parameters such as the maximum length of captions.
Define functions for preprocessing images, initializing hidden and cell states, generating captions, and generating stories. Preprocess the input image by resizing, flattening, and normalizing it. Initialize the hidden and cell states for the decoder LSTM. Generate captions for the input image using the encoder-decoder architecture. Generate stories by combining multiple captions generated for the same image.
Example usage: provide an image path and print the generated caption and story.

Load Trained Models:
Load the encoder and decoder models from their respective saved files.

Import Libraries:
Import necessary libraries such as pandas, NumPy, TensorFlow, etc.
Load Captions:
Load the captions from the dataset using pandas Data Frame.

Tokenize Captions:
Tokenize the captions to create a vocabulary and map each word to an index.
Save the tokenizer as a JSON file for future use.
Define Parameters:
Define parameters such as the maximum length of captions.
Define Functions:
Define functions for pre-processing images, initializing hidden and cell states, generating captions, and generating stories.
Pre-process Image:
Pre-process the input image by resizing, flattening, and normalizing it.
Initialize States:
Initialize the hidden and cell states for the decoder LSTM.
Generate Captions and stories:
Extract image features using the encoder model. Use the decoder model to generate words sequentially until the '<end>' token is encountered or the maximum length is reached. Generate stories by combining multiple captions generated for the same image.

Experiment and Result analysis:

A. SETUP AND CONFIGURATION

The deep learning models for this study were trained using the NVIDIA DIGITS toolbox on a Windows environment hosted on the Kaggle platform. The system configuration included an Intel Core i7 processor, 32GB of RAM, and four NVIDIA GeForce RTX 3080 GPUs.

B. RESULTS OF IMAGE CAPTION GENERATION

For training the caption generation model, 80% of the conceptual caption's dataset was randomly selected, while the remaining 20% was reserved for the validation purpose. The loss for the training and validation process will be given here, along with the loss, the BLUE scores for the various models that we have used will also be mentioned in detailed way.

S.no	Model	Loss
1.	VGG 16	3.8241
2.	Google Net(Inception)	4.2104
3.	Alex Net	4.6513

Table1: Represents the loss, BLEU-1 and BLEU-2 scores generated when the models vgg-16, Google Net and Alex Net are used

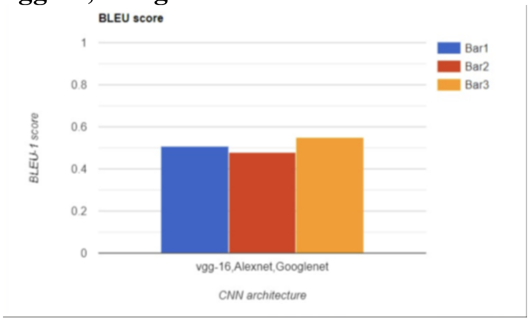

Figure 3: BLEU score obtained by using the models- Vgg-16, AlexNet and GoogleNet

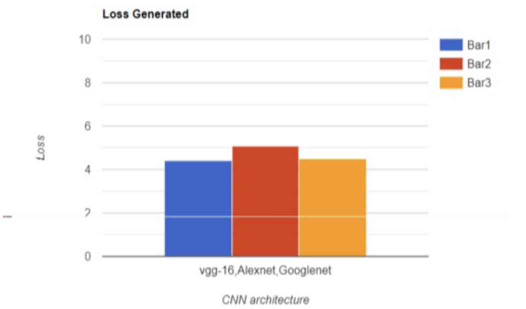

Figure 4: Loss generated by using the models- Vgg-16, AlexNet and GoogleNet

RESULTS OF STORY GENERATION

Figure 5 illustrates four diverse image captions and stories generated by the model, spanning genres like comics, horror, romance, and more. Utilizing deep visual-semantic alignment, the model crafts narratives that align with the image context. These narratives are not only semantically coherent but also tailored to various image styles, demonstrating the model's versatility in storytelling.

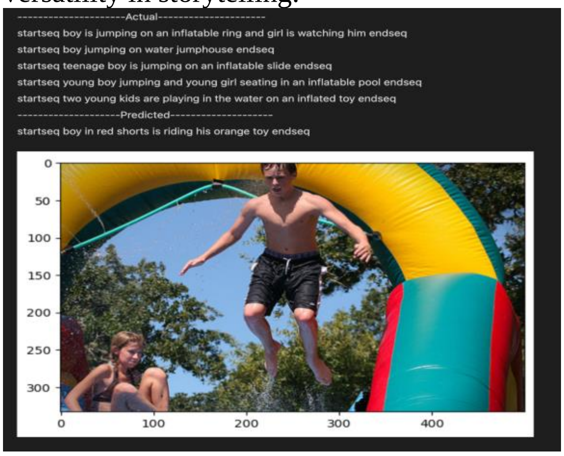

Figure 5 : Represents the image caption and story generated for the given image

In Figure 6 and Figure 7 we can compare the difference in the intuitive story that is generated when using encoders and decoders with LSTM and GPT-3 models.

Generated Story:

sheepdogs doorstep doorstep doorstep doorstep sheepdogs sheepdogs sheepdogs sheepdogs bares bares bares bares sideline sideline driver driver driver driver driver driver driver barking driver driver Boys Boys Boys adults tipped woman gith aided unoccupied bunch intensely tv march Trumpet Girls Girls brides 08 lifevests playtoy appearing lightly lightly inspects step sheepdogs doorstep doorstep doorstep doorstep sheepdogs sheepdogs sheepdogs sheepdogs bares bares bares bares sideline sideline driver driver driver driver driver driver driver barking driver driver Boys Boys Boys adults tipped woman gith aided unoccupied bunch intensely tv march Trumpet Girls Girls brides 08 lifevests playtoy appearing lightly lightly inspects step sheepdogs doorstep doorstep doorstep doorstep sheepdogs sheepdogs sheepdogs sheepdogs bares bares bares bares sideline sideline driver driver driver driver driver driver driver barking driver driver Boys Boys Boys adults tipped woman gith aided unoccupied bunch intensely tv march Trumpet Girls Girls brides 08 lifevests playtoy appearing lightly lightly inspects step

Figure 6: story generated by using encoders and decoders with LSTM models

Generated Story:

On a quaint countryside doorstep, a pack of eager sheepdogs await their master's return. Their relentless barking echoes through the serene landscape, drawing attention from nearby bystanders. Meanwhile, a diligent driver manoeuvres through the winding roads, accompanied by the distant sound of a trumpet's melody. Amidst the commotion, a group of boys playfully tip their hats to passing adults, while a woman, aided by her loyal sheepdogs, tends to her unoccupied flock. As the sun sets, the scene transforms into a bustling march, with girls adorned as brides and life vests appearing lightly on the horizon, each step inspected by the watchful eyes of the sheepdogs.

Figure 7: story generated by using GPT-3 model

Conclusion:

While AI programs excel in art and music, achieving proficiency in writing remains challenging due to the complexity of language. To facilitate story creation, a diverse dataset of romance and horror novels was collected. Overcoming language complexities, structural

variations, and vocabulary nuances poses significant challenges in this endeavor. For instance incorporating modules for grammar and contextual details could improve the accuracy of stories while allowing for narratives without sacrificing coherence. Furthermore utilizing a CNN model with the GRU technique can improvize the results originally needed for image processing. Developing an machine learning model specifically tailored for story creation and integrating it with NLP techniques for data processing shows potential for building an advanced story generator, in the future.

References:

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, ‘‘Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6077–6086.

[2] Bo Dai, Sanja Fidler, Raquel Urtasun, Dahua Lin, ‘‘Diverse Image Captioning via Group-Wise Discriminability,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2970–2979.

[3] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, ‘‘Tell Me What You See and I Will Show You Where It Is,’’ in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5867–5876.

[4] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Alan L. Yuille, ‘‘Image Captioning with Semantic Attention,’’ in Proc. IEEE Int. Conf. Comput. Vis., 2016, pp. 4651–4659.

[5] Y. Zhu et al., "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 19–27.

[6] Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher, ‘‘Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning,’’ in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 3242–3250.

[7] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille, ‘‘Attend to You: Personalized Image Captioning with Context Sequence Memory Networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10277–10286.

[8] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler, ‘‘Neural Storyteller: A Generative Model for Storylines,’’ in Proc. Conf. Neural Inf. Process. Syst., 2015, pp. 3104–3112.

[9] S. J. Rennie et al., "SCST: Reinforcement Learning of Sequence Generation Models without Likelihoods," in Proc. Conf. Neural Inf. Process. Syst., 2017, pp. 4565–4574.

[10] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2019, pp. 4171–4186.

[11] T.-H. K. Huang et al., "Visual Storytelling," in Proc. IEEE Int. Conf. Comput. Vis., 2016, pp. 1233–1242.

[12] L. Yao et al., "Image Captioning with Deep Bidirectional LSTMs," in Proc. Conf. Neural Inf. Process. Syst., 2017, pp. 104–113.

[13] A. Karpathy, L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3128–3137.

[14] J. Donahue et al., "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 2625–2634.

[15] Z. Yu et al., "Stacked Cross Attention for Image-Text Matching," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6666–6675.

[16] L. H. Li et al., "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training," in Proc. Conf. Neural Inf. Process. Syst., 2019, pp. 10462–10473.