World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Multimodal AI framework for image captioning, story generation and natural speech narration

Ashwani Attri, Priyanka Gudeboyena, Vaishnavi Chigurla *, Soumika Moluguri and Nithin Kasoju

Department of Computer Science and Engineering (Data Science), Ashwani Attri, ACE Engineering College, Telangana, India.

World Journal of Advanced Research and Reviews, 2025, 26(02), 1037-1044

Publication history: Received on 27 March 2025; revised on 03 May 2025; accepted on 06 May 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.2.1685

Abstract

With the increasing ubiquity of digital imagery, there is a growing need for intelligent systems capable of understanding visual content and expressing that understanding in human-like language. This paper presents a comprehensive AI-based pipeline that not only generates captions from images but also constructs vivid stories based on those captions and finally delivers them in a human voice. The proposed system integrates multiple components: a Convolutional Neural Network (VGG16) for extracting visual features, an LSTM-based sequence model for caption generation, GPT-2 for creative story generation, and Google Text-to-Speech (gTTS) for voice synthesis. The result is a multi-modal AI framework capable of transforming static images into rich, spoken narratives. This approach has applications in assistive technologies, interactive storytelling, content automation, and education. The proposed model is trained and evaluated on the Flickr8k dataset, demonstrating a viable path for automated visual storytelling.

Keywords: Image Captioning; CNN-LSTM; VGG16; GPT-2; Text-to-Speech (gTTS); Image-to-Story Generation; Natural Language Processing (NLP)

1. Introduction

The synergy between computer vision and natural language processing has led to groundbreaking innovations in artificial intelligence. Deep learning techniques have allowed machines to interpret complex patterns in both images and text, giving rise to applications like image captioning, automated storytelling, and human-computer interaction systems. However, while each of these areas has matured in isolation, combining them into a seamless, human-centric experience remains a frontier in AI research.

Humans possess a remarkable ability to perceive a visual scene, articulate its content in descriptive language, and even extrapolate imaginative stories from it. For instance, when looking at a photograph of a child playing with a puppy in a park, a human observer might not only say, “A child is playing with a dog,” but also create an engaging narrative such as, “On a sunny afternoon, Emma found a new best friend in the park.” Emulating this level of interpretation requires more than just object detection or sentence generation—it calls for contextual understanding, narrative imagination, and voice delivery.

This paper proposes a system that emulates this human storytelling process. It starts by analyzing an image using a pre-trained VGG16 CNN model to extract high-level features. These features are then fed into an LSTM-based decoder, trained on the Flickr8k dataset, to generate a concise caption. Next, a transformer-based language model, GPT-2, takes the caption and expands it into a creative short story. Finally, the story is converted into natural-sounding speech using gTTS, creating a fully immersive and interactive storytelling experience.

[*] Corresponding author: Ch Vaishnavi

The integration of image understanding, natural language generation, and voice synthesis provides a powerful tool for accessibility—especially for the visually impaired—where the ability to "see through words" becomes transformative. The system also opens up new avenues in education, interactive media, and digital content creation.

2. Literature Review

The domain of image captioning has evolved through several phases. Early methods relied on template-based approaches and rule-based systems that had limited generalization capability. With the advent of deep learning, encoder-decoder architectures became the norm. Vinyals et al. (2015) introduced the "Show and Tell" model, a breakthrough in combining CNNs and RNNs for end-to-end caption generation. Later models such as "Show, Attend and Tell" incorporated attention mechanisms, improving focus on relevant parts of the image during word generation.

Story generation, on the other hand, has seen progress through large-scale transformer-based models. Radford et al.'s GPT-2 (2019) demonstrated the ability to generate fluent, contextually coherent text, opening the door for creative applications such as dialogue agents, automatic writers, and storytelling bots. However, these models operate purely in the text domain, and without visual grounding, their stories can be contextually generic or misaligned with visual prompts.

Text-to-speech synthesis has also significantly improved. Google's gTTS and other neural TTS engines like Tacotron and WaveNet have made it possible to generate natural-sounding speech with minimal latency. These tools provide the auditory interface for AI systems, especially in accessibility and human-computer interaction contexts.

Despite these individual advancements, very few systems have attempted to integrate visual understanding, language generation, and speech synthesis into a single pipeline. This research builds upon these foundational works to propose an end-to-end model for automated image-based storytelling.

3. Existing System

Several standalone systems exist that perform well in isolation:

3.1. Image Captioning Systems

Deep learning models like "Show and Tell" and "Neuraltalk2" generate concise captions for images, focusing on object recognition and sentence fluency. These systems, however, are limited to short phrases and lack narrative capability.

3.2. Story Generation Tools

GPT-2 and its successors (e.g., GPT-3, GPT-4) have revolutionized text generation, producing creative and engaging content. Yet, these tools require carefully crafted prompts and do not accept visual inputs directly.

3.3. Text-to-Speech Engines

Tools such as Google TTS and Amazon Polly provide high-quality speech output. They are widely used in accessibility applications, virtual assistants, and audio content generation.

Each of these systems is useful independently, but the lack of integration creates friction when building an application that seeks to mimic human storytelling from visual stimuli. The need for a unified, automated, and contextually coherent system remains largely unmet.

4. Proposed Model

To address the fragmentation in existing solutions, we propose a unified multi-modal architecture that emulates a human-like storytelling process from visual input. The proposed model consists of four tightly integrated modules:

4.1. Visual Feature Extractor

- Uses the pre-trained VGG16 model to extract high-level features from input images.
- Outputs a 4096-dimensional feature vector representing visual semantics.

4.2. Caption Generator

- Employs a Tokenizer and LSTM-based decoder.
- Trained on the Flickr8k dataset to convert visual features into grammatically correct and contextually relevant captions.

4.3. Story Generator

- Utilizes GPT-2 for text generation.
- Takes the generated caption as a prompt and produces a short, creative story based on it.

4.4. Text-to-Speech Synthesizer

- Converts the story into an English voice using gTTS.
- Provides auditory feedback for accessibility and engagement.

Together, these components form an end-to-end pipeline capable of transforming static visual inputs into rich, spoken narratives.

5. Methodology

The proposed system follows a modular deep learning-based methodology to transform static images into rich, spoken narratives. First, input images are preprocessed and passed through a pre-trained VGG16 model to extract deep visual features. These features are then fed into a CNN-LSTM network that generates a meaningful caption word by word. The caption is used as a prompt for GPT-2, which generates a vivid, coherent story capturing the scene's context. Finally, the story is converted into speech using Google Text-to-Speech (gTTS), completing the visual-to-audio transformation pipeline.

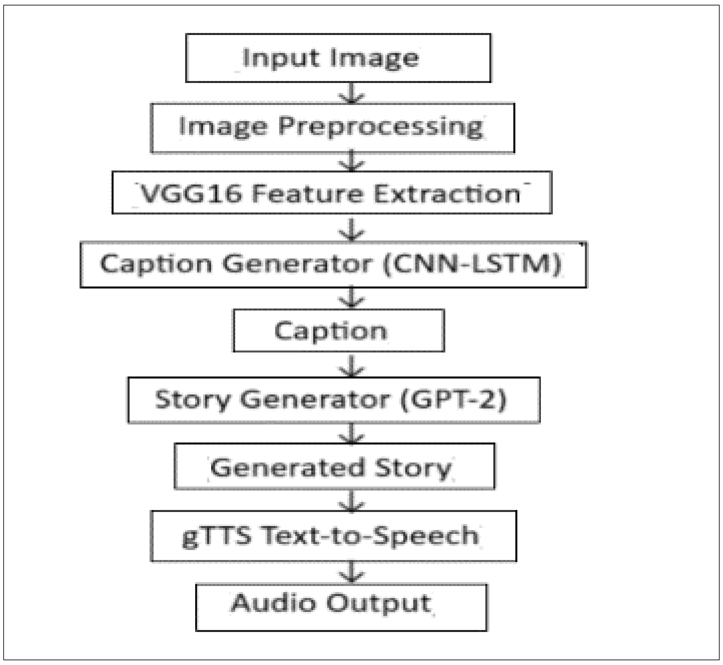


Figure 1 Methodology(Source: Authors)

5.1. System Architecture

The architecture of the proposed system integrates Computer Vision, Natural Language Processing, and Text-to-Speech (TTS) in a multi-stage pipeline that automates storytelling from visual data. The entire pipeline is modular and follows a clear sequence: image preprocessing, feature extraction, caption generation, story generation, and speech synthesis. Below is an overview and breakdown of each major component:

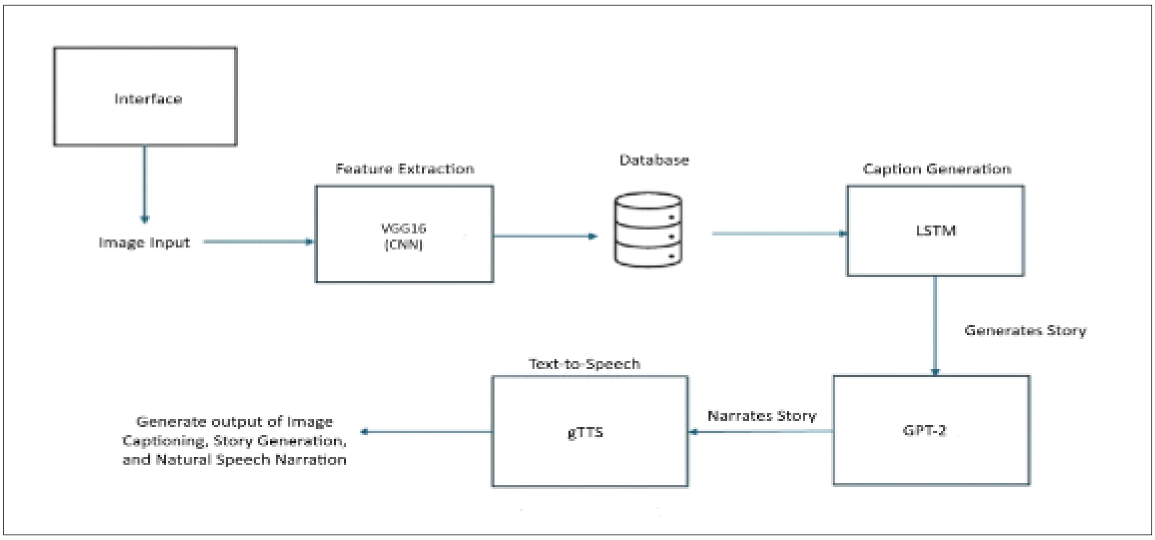


Figure 2 System Architecture(Source: Authors)

5.2. Image Input Module

The first step in the system is to handle the image input. Users either provide an image manually or select one from a dataset, like the Flickr8k dataset. The input image is typically in various formats and sizes, so this module standardizes the image by resizing it to a fixed size (224x224 pixels). Additionally, the image is converted into an RGB format, ensuring it is compatible with the VGG16 model for feature extraction. The input module is crucial for maintaining consistency across all image data used in the subsequent stages.

Once the image is processed, it is transformed into a NumPy array for easy manipulation in TensorFlow or Keras. This module is designed to support various image formats and can be extended for real-time image upload, making it more user-friendly for different applications.

5.3. Feature Extraction Module (CNN - VGG16)

After the image is preprocessed, the Feature Extraction Module utilizes a pre-trained VGG16 model to extract high-level visual features. VGG16 is a convolutional neural network (CNN) model that was originally trained for image classification tasks on large datasets such as ImageNet. In this system, instead of using the model for classification, we utilize the penultimate layer of the network to capture the rich features of the image.

The model processes the input image and generates a 4096-dimensional feature vector. This vector represents the most salient features of the image, such as objects, textures, and spatial relationships, which are crucial for generating an accurate description. These features serve as a compressed yet informative representation of the image, capturing the essential visual information that is passed to the caption generation module.

5.4. Caption Generation Module (CNN-LSTM)

The Caption Generation Module is responsible for generating a textual description of the image. This module combines the visual features extracted by the CNN (VGG16) with a sequence of words to generate coherent and descriptive captions. The architecture employed in this system is a combination of CNN and LSTM (Long Short-Term Memory), which has proven effective for sequence generation tasks.

The CNN extracts the visual features, and these features are fed into the LSTM, which is trained to predict a sequence of words. The LSTM is an advanced type of recurrent neural network (RNN) that excels at capturing temporal dependencies in data, making it ideal for sequence prediction tasks like captioning. The Embedding Layer transforms words into dense vector representations, allowing the LSTM to process these word sequences effectively.

The system generates the caption word by word, starting with a token like startseq and iteratively predicting the next word based on the previous ones. The process continues until an endseq token is predicted, signaling the end of the caption.

5.5. Text Enhancement Module (Story Generator using GPT-2)

Once the caption is generated, the Text Enhancement Module uses a pre-trained GPT-2 model to transform the short caption into a full story. This is where the system's creativity is injected. GPT-2, a powerful transformer-based language model, is designed for text generation and can create highly coherent and contextually relevant text based on a given prompt. In this case, the caption serves as the prompt.

The module sends the caption to GPT-2, which then generates a detailed, vivid story. The story goes beyond simple captioning, weaving a narrative that adds depth and context to the image. For example, a caption like “a dog running in the park” might evolve into a story describing the dog’s adventure in the park, its interaction with other animals, and its surroundings.

The story generation is parameterized by settings such as temperature (which controls the randomness of the output), top-p (which regulates the diversity of the predictions), and max_length (which limits the number of generated words). These settings are optimized to produce stories that are both creative and coherent.

5.6. Text-to-Speech Module (gTTS)

The generated story is passed to the Text-to-Speech (TTS) Module, where it is converted into spoken words. This module uses Google's Text-to-Speech API (gTTS), which is a lightweight and easy-to-use tool for converting text into speech. gTTS supports multiple languages, but in this case, the system generates English audio.

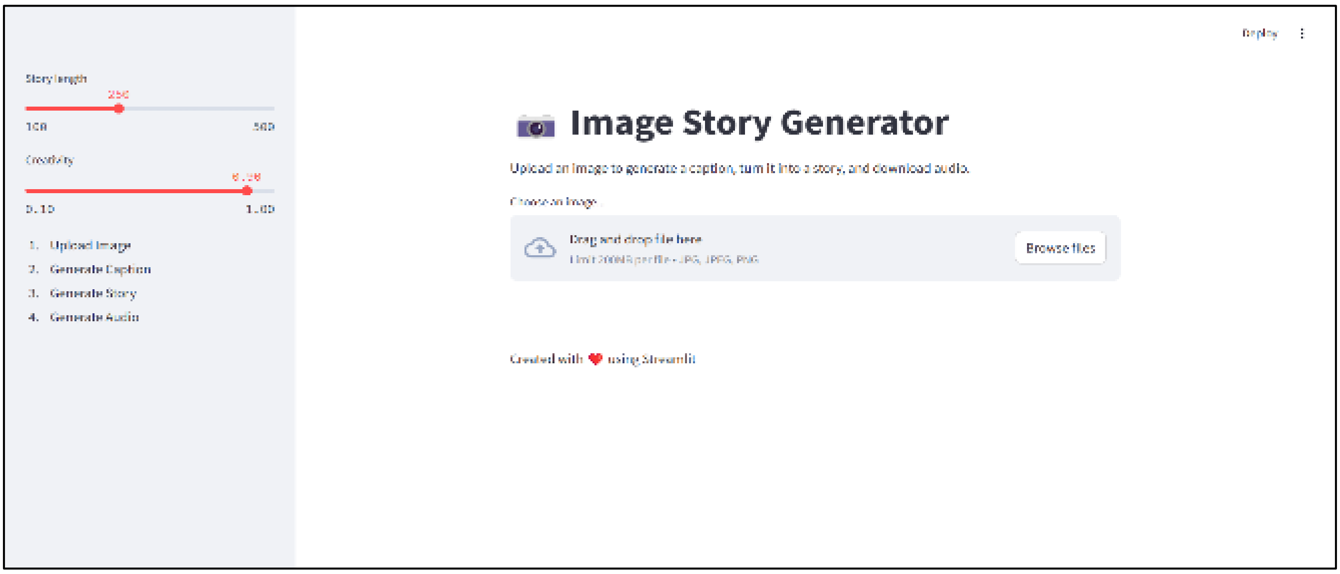
Once the story is converted into an audio file, the gTTS module saves it as an MP3 file. The MP3 file can be played back for the user, providing an immersive audio experience. This module is particularly useful for creating accessible content, enabling the visually impaired to enjoy the generated stories through audio.

The speech synthesis is not only a critical component for accessibility but also adds a dynamic, lifelike layer to the system, making the narrative experience more engaging.

5.7. Output Module

The Output Module is responsible for playing the generated audio file and displaying relevant visual information. In this module, the audio file generated by the TTS system is played back using the playsound library, allowing the user to listen to the story.

6. Results and Discussion



Story length: 250 (100–500)

Creativity: 0.90 (0.10–1.00)

1. Upload Image
2. Generate Caption
3. Generate Story
4. Generate Audio

Image Story Generator

Upload an image to generate a caption, turn it into a story, and download audio.

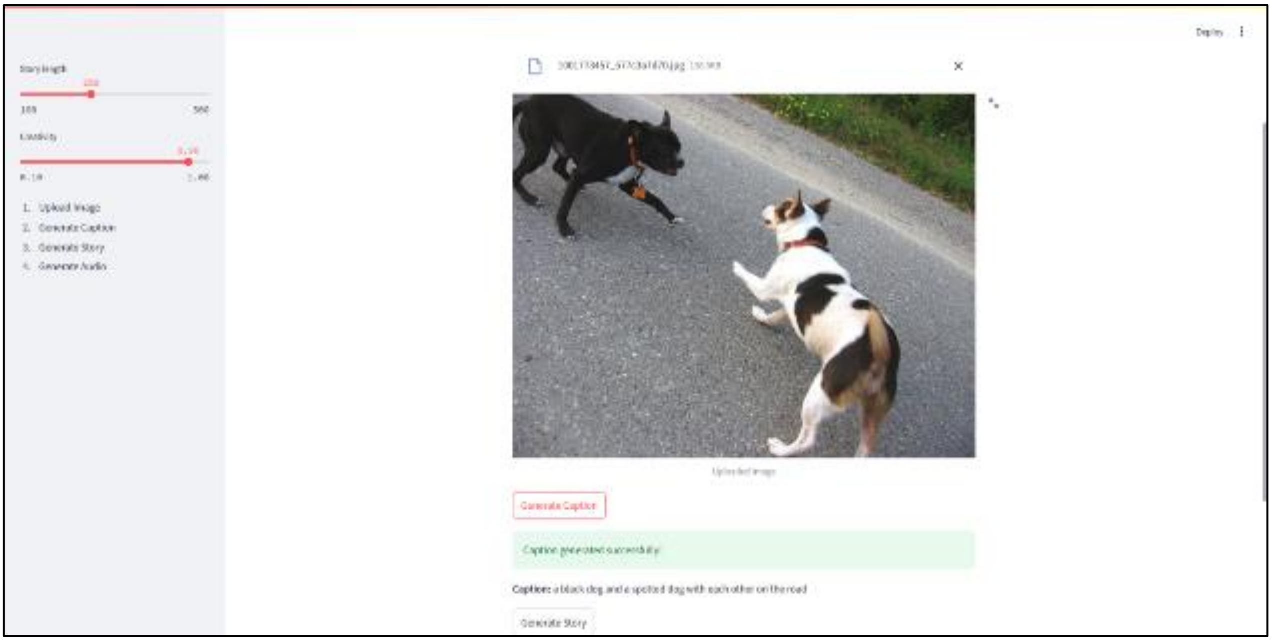
Choose an image

Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG

Created with ❤ using Streamlit

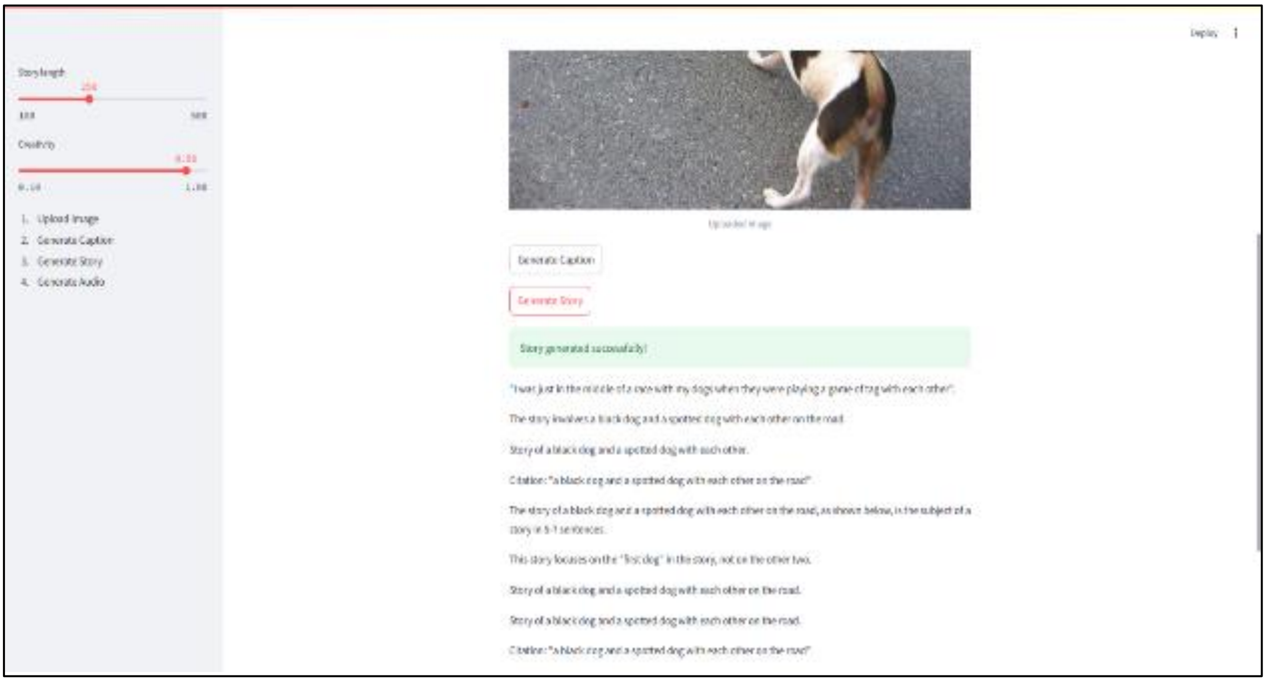
Figure 3 User Interface(Source: Authors)



Caption generated successfully!

Caption: a black dog and a spotted dog with each other on the road

Figure 4 Input Image(Source: Authors)



Story generated successfully!

"I was just in the middle of a race with my dogs when they were playing a game of tag with each other".

The story involves a black dog and a spotted dog with each other on the road.

Story of a black dog and a spotted dog with each other.

Citation: "a black dog and a spotted dog with each other on the road"

The story of a black dog and a spotted dog with each other on the road, as shown below, is the subject of a story in 5-7 sentences.

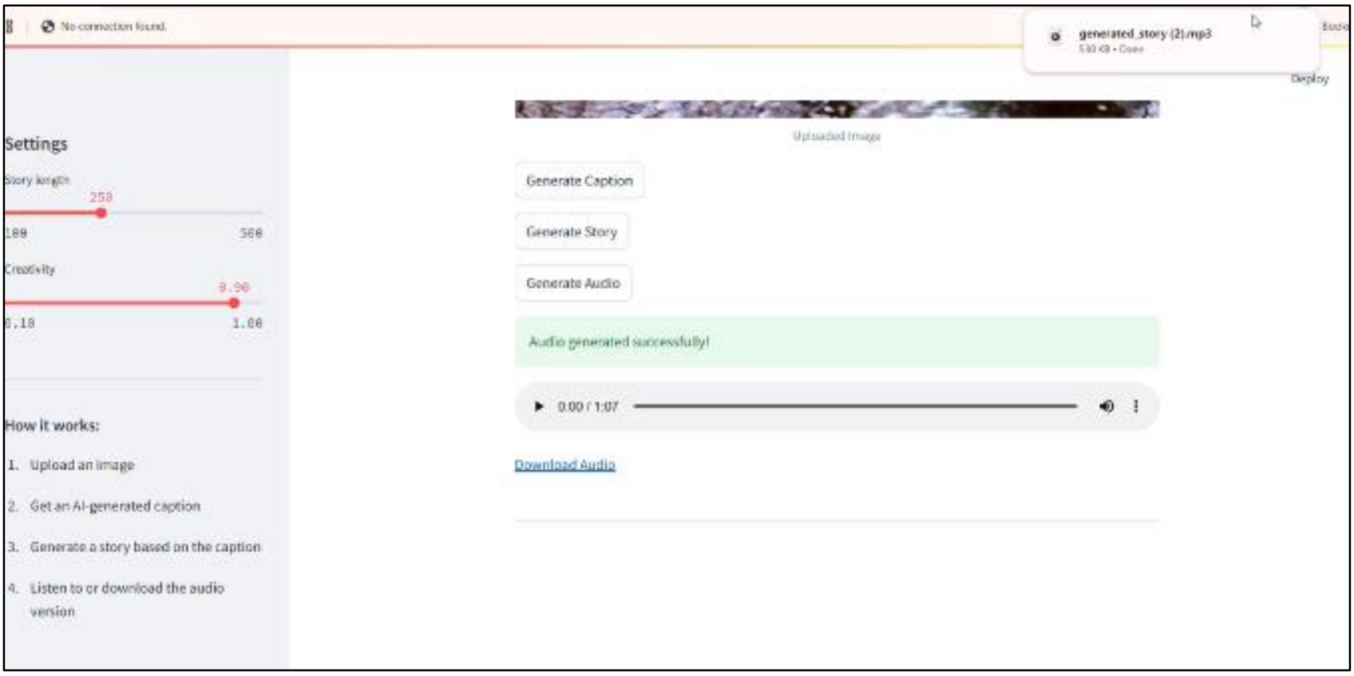
This story focuses on the "first dog" in the story, not on the other two.

Story of a black dog and a spotted dog with each other on the road.

Story of a black dog and a spotted dog with each other on the road.

Citation: "a black dog and a spotted dog with each other on the road"

Figure 5 Caption and Story generation(Source: Authors)



No connection found.

generated_story (2).mp3
530 KB • Done

Settings

Story length
250
100 500

Creativity
0.90
0.10 1.00

How it works:

1. Upload an image
2. Get an AI-generated caption
3. Generate a story based on the caption
4. Listen to or download the audio version

Audio generated successfully!

0:00 / 1:07

Figure 6 Audio file stored as .mp3(Source: Authors)

7. Conclusion

This paper presents an integrated AI system that performs image captioning, story generation, and text-to-speech conversion in a unified pipeline. By leveraging deep learning techniques in computer vision and natural language processing, the model offers a novel approach to creating engaging, human-like narratives from static images.

Beyond its technical implementation, the system humanizes machine perception—transforming passive image recognition into an active storytelling experience. Applications of this work include educational tools, assistive technologies for the visually impaired, interactive content platforms, and creative writing aids.

Future work will explore enhancements such as attention mechanisms, multimodal transformers, multilingual support, and real-time web or mobile deployment.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 2048–2057).http://proceedings.mlr.press/v37/xuc15.html

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI.models/language_models_are_unsupervised_multitask_learners.pdf

[3] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and Tell: A Neural Image Caption Generator*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3156–3164).https://doi.org/10.1109/CVPR.2015.7298935

[4] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., et al. (2017). *Tacotron: Towards End-to-End Speech Synthesis*. In *Proceedings of Interspeech 2017*.https://doi.org/10.21437/Interspeech.2017-1452

[5] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv preprint arXiv:1609.03499.https://arxiv.org/abs/1609.03499

Author’s short biography

Mr. Ashwani Attri: Mr. Ashwani Attri, Completed his B.Tech and M.Tech in CSE from IIT Kharagpur, He worked in IT Sector as Software Engineer and is Currently working as Assistant Professor, Department of CSE(Data Science), ACE Engineering College. He aim to inspire students and contribute to advancements in technology through his work.	
Priyanka Gundeboyena: Priyanka, a final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). Her academic journey has been shaped by a strong commitment to exploring data-driven solutions for real-world challenges. Over the course of my studies, She have developed a profound interest in areas such as cybersecurity, machine learning and networking.	
Vaishnavi Chigurla: I am Vaishnavi, a final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). My academic journey has led me to explore the dynamic fields of cybersecurity, machine learning, artificial intelligence, and networking. With a keen interest in innovation, I strive to bridge the gap between data science and real-world applications, continuously expanding my knowledge and expertise.	
Soumika Moluguri: Soumika, a final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). Her academic journey has led me to explore diverse domains, with a particular focus on web development, machine learning, and artificial intelligence. Skilled in building dynamic and efficient web solutions, She is committed to continuous learning and innovation to bridge the gap between data science and modern web technologies.	
Nithin Kasoju: Nithin, a final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). His academic journey has been shaped by a strong commitment to exploring data-driven solutions for real-world challenges. Over the course of his studies, He have developed a profound interest in areas such as machine learning, artificial intelligence, and networking.	