GenStory: Automated Generation of Children's Stories with Multimodal AI

1st Chavez Piguave Aldrin Josue
ECMC
Yachay Tech
Urcuquì, Ecuador
aldrin.chavez@yachaytech.edu.ec

2nd Sanchez Salazar Kevin Josue
ECMC
Yachay Tech
Urcuqui, Ecuador
kevin.sanchez@yachaytech.edu.ec

3rd Peñaherrera Loor Jhony Alfonso
ECMC
Yachay Tech
Urcuquí, Ecuador
jhony.penaherrera@yachaytech.edu.ec

4th Chalacán Sandoval Mateo Alejandro
ECMC
Yachay Tech
Urcuquí, Ecuador
mateo.chalacan@yachaytech.edu.ec

***Abstract*—This project presents the design and development of a multimodal mobile application that automatically generates personalized children's stories by integrating image and text using deep learning models. Driven by the increasing use of mobile devices among children as passive entertainment, the application offers an educational and creative alternative that encourages reading and imagination. The system combines a pre-trained ResNet18 model (tuned on Places365) for scene classification with the GPT-2 language model for narrative generation. By allowing users to upload an image and provide a brief textual description, the application generates coherent and engaging stories in English, adapted to the visual and semantic context. The interface, developed with Streamlit, provides an accessible and interactive narrative experience, suitable for children with basic reading skills. Evaluation results demonstrate the system's ability to produce fluent and imaginative narratives, aligned with both visual information and user prompts. Despite certain limitations in terms of narrative structure and dataset specificity, the project demonstrates the potential of multimodal AI for educational purposes. Future improvements include fine-tuning models on domain-specific datasets, improving narrative consistency for children, and expanding multilingual support.**

***Index Terms*—Multimodal AI, Story Generation, Children's Education, Deep Learning, ResNet18, GPT-4, Human-AI Interaction, Narrative Generation, Educational Technology, Computer Vision and NLP.**

I. Introduction

The evolution and advancement of artificial intelligence-based technologies have opened up new possibilities for the innovative creation of highly personalized content, which in turn enables the development of much more interactive experiences tailored to the diverse needs and preferences of different audiences. In this context, the project in question proposes the development of a multimodal mobile application specifically designed for the automatic generation of children's stories, combining visual and textual recognition to produce original narratives specifically designed to captivate children's imaginations.

The main objective of this proposal is to integrate artificial intelligence tools for recreational and educational purposes, thus contributing to greater promotion of reading and the development of children's imagination. This idea arises from the current concern about the increasing use of mobile devices by children, driven primarily by recreational activities related to video games, which, in many cases, has led to the unfortunate displacement of traditional activities such as reading or narrative creation. (Binford, 2015)

This application directly addresses this issue by offering an alternative solution that strategically leverages technology to inspire and motivate children to actively participate in the creative process of storytelling, giving them the unique opportunity to select a setting through a visual image and enrich it with a concise textual description that includes characters, conflicts, or other elements essential to the development of the plot. The app is designed for children with basic reading skills and is intended for independent use or with adult guidance. Its primary purpose is twofold: recreational and educational, offering children the opportunity to develop reading-related skills, narrative comprehension, and greater creativity. (Elsayad, 2014)

From a technical perspective, the system is based on two main components: an image classification model based on the ResNet architecture (specifically, using resnet18 pre-trained with Places365), whose task is to accurately identify the visual scene provided by the user; and a text generation model based on GPT-4, designed to produce a message-based narrative, constructed using the detected visual category along with the descriptive text entered by the user. The synergy between these two models facilitates the generation of coherent and contextually relevant stories in English, thus expanding its potential for international application and use. This report details the design, implementation, and testing process of the application, as well as the various methodological and technical considerations adopted throughout its development process.

II. Methodology

Operational Flow and Technical Pipeline

The operational flow of our application revolves around the integration of two pre-trained deep neural network architectures, enabled through an interactive interface built with `Streamlit`. This framework not only facilitates user image upload and collection of textual descriptions, but also allows for transparent visualization of the multimodal processing underlying the backend.

App Development and Technical Pipeline

Upon uploading an image, the system invokes the **ResNet18** architecture, an 18-layer residual convolutional network known for its ability to mitigate the problem of gradient vanishing through skip connections. This specific variant has been trained on the **Places365** dataset, which covers 365 categories of natural and human-made scenes, from beaches and forests to urban interiors. The model is downloaded and initialized in the environment, ensuring exact matching of weights and architecture (including the number of filters, convolutional layers, residual blocks, and ReLU activation functions).

Image preprocessing is essential: it is normalized and adjusted to the dimensions required by the network, ensuring compatibility with the pre-trained weights. Once passed through the network, the image undergoes a series of convolutional and pooling operations, followed by fully connected layers. The output vector, of dimension 365, represents the probability of belonging to each scene class. Here, the **softmax** function transforms the logits into an interpretable probability distribution, from which the class with the highest score is selected as the predominant scenario.

On the other hand, narrative generation relies on the **GPT-4** model, a transformer architecture that has revolutionized natural language processing. GPT-4 is composed of multiple autoregressive attention layers, allowing it to model long-term dependencies and generate coherent and contextualized text. The prompt that feeds GPT-4 is not arbitrary: it is constructed by integrating the scene label predicted by ResNet18 along with the description provided by the user, following the template:

> *"The story takes place near a {scene}, where {user_input}."*

This prompt design is key, as it conditions text generation to a visual-semantic context, thus achieving synergy between computer vision and natural language processing.

It should be noted that the choice of pre-trained models not only optimizes computational resources but also allows leveraging the knowledge acquired from large corpora of data, both visual and textual. However, it is important to recognize that the generalization of these models may be limited by biases inherent to the original datasets, and that the narrative coherence generated by GPT-4, although surprising, is not free from semantic or stylistic inconsistencies.

In short, the app not only automates scenario classification and story generation, but also exemplifies the potential of multimodal artificial intelligence to create interactive and personalized experiences, integrating advanced deep learning techniques into an end-user-friendly workflow.

III. System Implementation

The system is divided into two main components: a backend server responsible for scene recognition and story generation, and a mobile frontend application designed for user interaction and input collection. Both components are integrated to deliver a seamless storytelling experience.

A. Backend Implementation

The backend was developed using **FastAPI**, exposing an HTTP endpoint for story generation. It is composed of several key modules:

- `main.py`: Defines the API endpoint `/generate`, which receives a user-uploaded image and a prompt. It preprocesses the image, invokes the ResNet18 model for scene classification, constructs a prompt combining the predicted scene and the user description, and calls the `GPT-4o-mini` model from OpenAI to generate the story.
- `model_utils.py`: Loads the pre-trained `resnet18_places365.pth.tar` model and its associated category labels from `categories_places365.txt`. It performs image normalization, resizing, and classification, returning the most probable scene and its confidence score.
- `schemas.py`: Defines the output data structure using Pydantic's `BaseModel`, ensuring consistency in API responses.

The backend is fully containerized using Docker, with a minimal image based on `python:3.11-slim-bullseye`. All dependencies are declared in a `requirements.txt` file, and environment variables such as the OpenAI API key are managed via a `.env` file. The container listens on port 8000 and exposes the FastAPI Swagger documentation interface for testing. The backend is temporarily exposed using **Ngrok** via the public URL: `https://urchin-star-cattle.ngrok-free.app`.

B. Frontend Implementation

The mobile application was developed using **React Native** with **Expo**, styled using NativeWind (Tailwind CSS for React Native). It provides an intuitive interface where users can:

- Upload or capture an image to define the story setting.
- Input a brief textual description of the characters.
- Submit both inputs to the backend and receive a story.
- View the predicted image scene, and read the story in real time.

The core component, `StoryCraftApp`, manages states for image URI, user input, and generated story. It uses the `expo-image-picker` library to acquire images and sends them via a `multipart/form-data` POST request to the backend. Upon receiving a response, it displays the story to

the user in a scrollable view. The interface mimics modern children's applications with animated icons, color gradients, rounded cards, and user-friendly typography.

The mobile app successfully integrates with the backend, forming a fully functional multimodal pipeline from image and text input to AI-generated story output.

IV. Frontend Integration

The mobile application was developed using **React Native** with `Expo` and styled with standard `StyleSheet` configurations to ensure lightweight performance. Key elements of the app include:

- **Image Selection**: The app allows users to either take a photo using the device's camera or upload one from the gallery using `expo-image-picker`.
- **Text Input**: A multiline `TextInput` field lets users enter a short story idea or character description.
- **Story Generation Flow**:
  - Once the user selects an image and writes a description, a request is sent to the backend endpoint `/generate`.
  - The app shows a loading indicator while the story is being generated.
  - When a response is received, the story is rendered in a scrollable view using the fetched data.
- **UI Aesthetics**: The app mimics a storytelling visual identity with themed icons (Ionicons), vibrant gradients, and child-friendly fonts and layouts to enhance usability and appeal.

This seamless front-back communication provides an engaging and intuitive storytelling experience tailored for young users and parents alike.

V. Results

To illustrate the functionality and output quality of the application, the following example demonstrates a complete interaction between the user and the system:

- **Input Image:**

- **User Description:**
  "A curious boy finds a silver bill hidden under a bench. That night, a driverless, illuminated train arrives, inviting him on a mysterious journey."
- **Detected Scene (ResNet18 Prediction):**
  `forest_path`
- **Generated Story (GPT-4 Output):**

  Once upon a time, there was a curious boy named Emiliano who lived in a small town surrounded by mountains. Despite his small size, Emiliano was known for his inquisitive nature and insatiable curiosity. Every day after school, he would head to the train station, a facility that had been abandoned for years but still retained its charm.

  One gray, overcast day, while exploring the old train station, Emiliano found a shiny silver bill hidden under a worn wooden bench. The bill seemed magical, with unfamiliar characters engraved on it, and it emitted an ethereal glow. Emiliano, fascinated, put it in his pocket and ran home.

  That night, as he lay in bed, Emiliano noticed the moonlight illuminating the bill in his pocket. He decided to go out to the train station, driven by curiosity and excitement. Upon his arrival, he was met with a surprising sight. A bright, luminescent train, seemingly made of stars and dreams, was parked on the platform.

This result exemplifies the system's ability to understand a visual scene, interpret a user's idea, and generate a coherent narrative that aligns with both inputs. Despite its general-purpose nature, the GPT-4 model provides a fluent and imaginative story, especially suitable for children.

VI. Conclusions and Future Work

The development of this multimodal application has demonstrated the feasibility and potential of integrating pre-trained deep learning models to generate personalized and contextually appropriate narratives for children. Through the combination of ResNet18 for scene recognition and GPT-4 for story generation, the system effectively translates a visual-textual input into a narrative output that is both engaging and imaginative.

However, several limitations were identified. Firstly, GPT-4, being a general-purpose language model, lacks a specialized narrative structure tailored for children, which occasionally results in inconsistencies in tone or style. Secondly, the use of a broad scene classification model like ResNet18 trained on the Places365 dataset, although versatile, may not align perfectly with scenes typically present in children's storybooks or environments.

For future iterations of this project, the following improvements are proposed:

- **Dataset Customization:** Development of a domain-specific dataset composed of illustrated children's scenes to fine-tune both the vision and language models.
- **Custom Language Model:** Training or fine-tuning a narrative generation model specialized in children's literature, incorporating age-appropriate vocabulary, structures, and themes.
- **User Feedback Loop:** Including interactive feedback mechanisms to allow users (children or guardians) to

influence the narrative direction or provide corrections, enhancing the interactivity and educational value.

- **Multilingual Support:** Expanding the application's language capabilities to support other native languages, potentially including minority languages, to promote cultural diversity and inclusion.

In conclusion, this project lays the foundation for a novel educational tool that leverages the creative capabilities of artificial intelligence. By continuing to refine the underlying models and adapt them to the target audience, future versions of this application could provide a transformative experience for young readers and storytellers alike.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, arXiv:1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114

[9] S. Liu, "Wi-Fi Energy Detection Testbed (12MTC)," 2023, gitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC

[10] "Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009." U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2

[11] K. Eves and J. Valasek, "Adaptive control for singularly perturbed systems examples," Code Ocean, Aug. 2023. [Online]. Available: https://codeocean.com/capsule/4989235/tree

[12] Binford, W. (2015). The Digital Child. Social Science Research Network. https://doi.org/10.2139/SSRN.2563874

[13] Elsayad, N. A. (2014). Children's playgrounds & Recreation Areas. IOSR Journal of Humanities and Social Science, 19(3), 45–53. https://doi.org/10.9790/0837-19314553