SNAPCAP

Image Caption Generation using VGG16 and LSTM

PROBLEM

The goal of this project is to build an AI model that can automatically generate meaningful captions for images, similar to how a human would describe them.

It combines Computer Vision (to understand image content) and Natural Language Processing (to generate text), making images more accessible and easier to categorize.

This technology is valuable in real-world areas such as:

- Helping visually impaired users understand images
- Improving photo organization in digital libraries
- Enhancing content tagging for social media and search engines

DATASET

To train and evaluate our model, we used the Flickr8k dataset, which contains:

- 8,000 images of everyday scenes
- 5 human-written captions per image
- Each caption describes the scene in simple English, providing diverse yet relevant sentence structures.

METHODS

Feature Extraction (VGG16)

A pre-trained VGG16 model extracts high-level visual features from images. These are used as input for the captioning model.

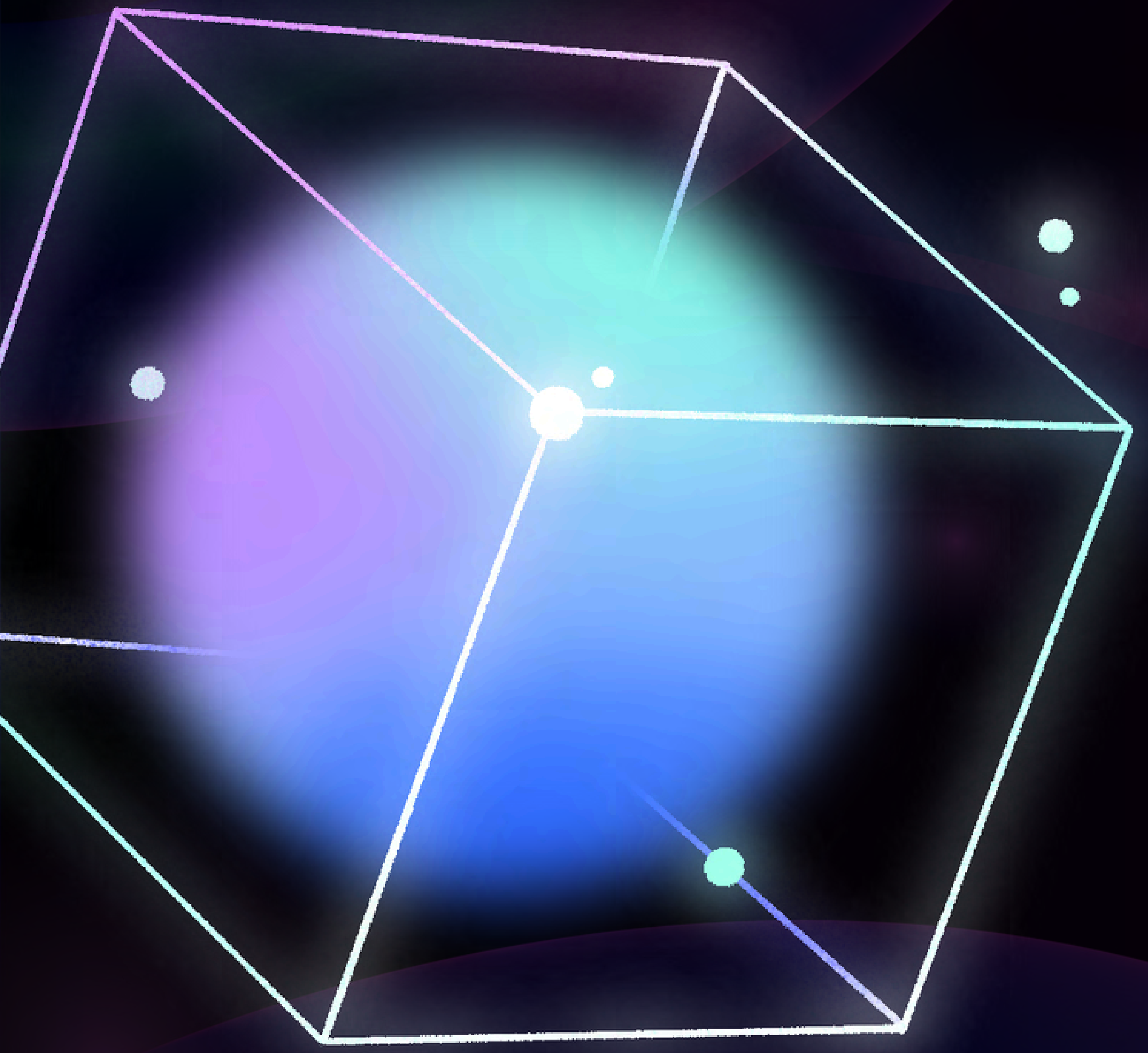
Caption Generation (LSTM Decoder)

An LSTM network receives the image features and generates a sequence of words to form the caption.

Training Approach

- Input: (Image features, partial caption)
- Output: Next word in the caption
- Optimizer: Adam
- Loss Function: Categorical Crossentropy

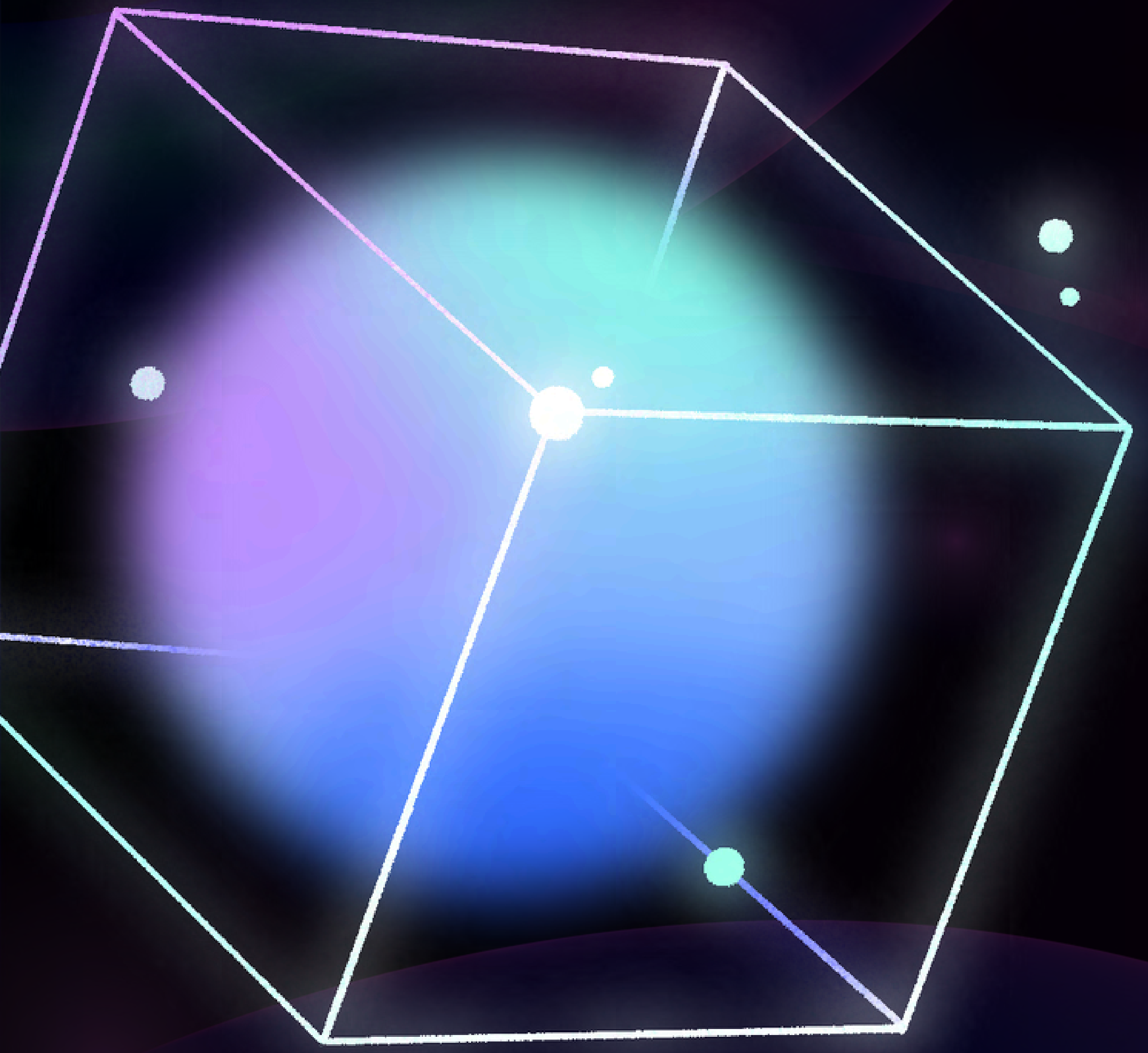
MODEL OBSERVATIONS



- The model is able to generate relevant and context-aware captions.
- BLEU score suggest the model performs well in aligning semantically with the actual captions.

Sample Output:

- Image: A dog running on grass
- Generated Caption: A black dog is running.



MODEL PERFORMANCE AND OBSERVATIONS

Key Findings:

- Good at recognizing simple objects and actions
- Struggles with complex sentence structures or unusual scenes
- Captions are often accurate but sometimes generic

THANK YOU!