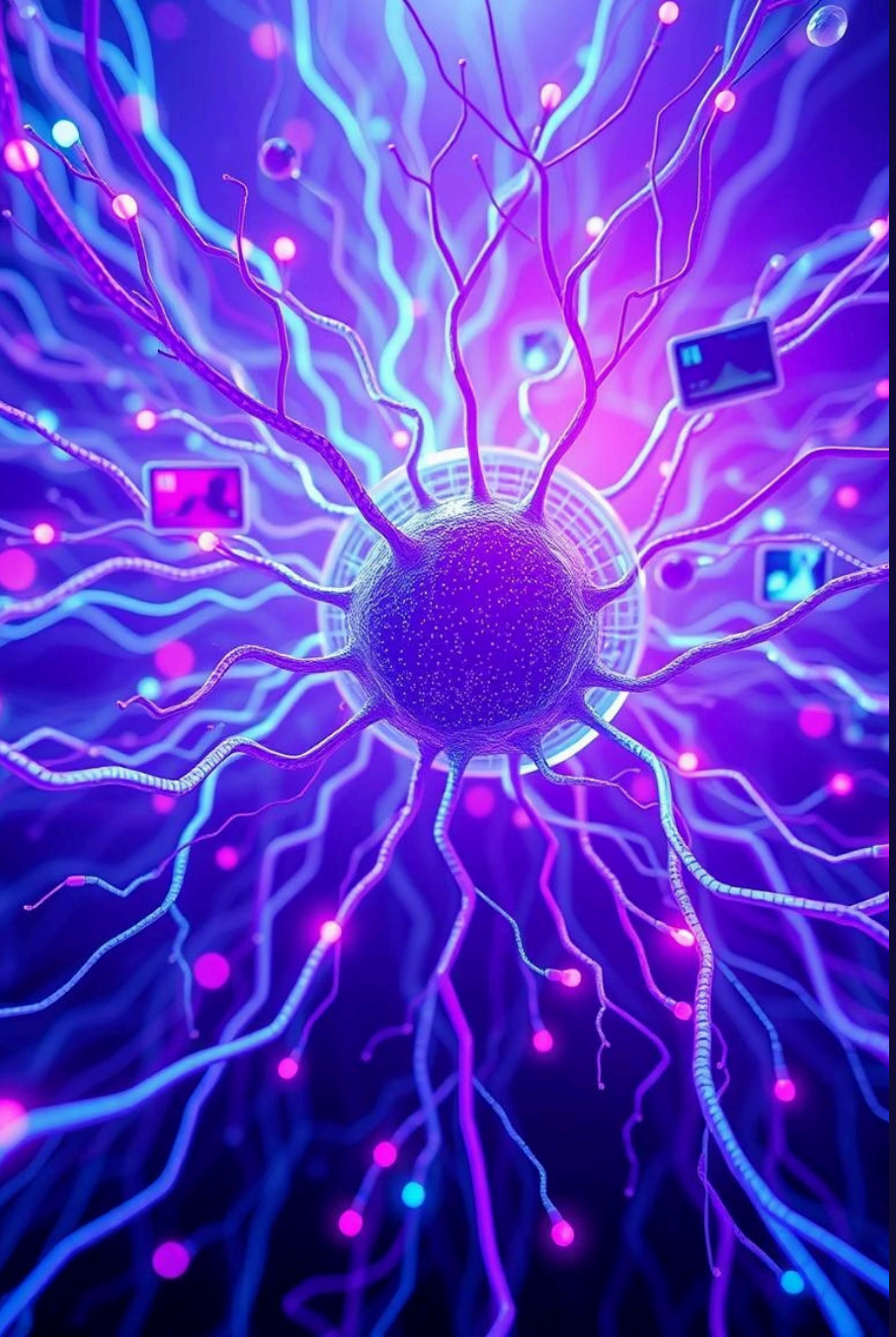


Deep Learning for Image and Video Captioning

The Image and Video Caption system employs advanced deep learning techniques to automatically generate rich, human-like descriptions. By integrating convolutional neural networks, sequence models, and transformers, it bridges vision and language domains. This presentation explores the methodology, evaluation metrics, and key findings, emphasizing recent advances in contextual understanding and temporal-spatial coherence in captions.

Team Members: Umangi Nigam E22CSEU0526, Tushar Swarnkar E22CSEU0436

Data Collection and Preparation

Datasets Used

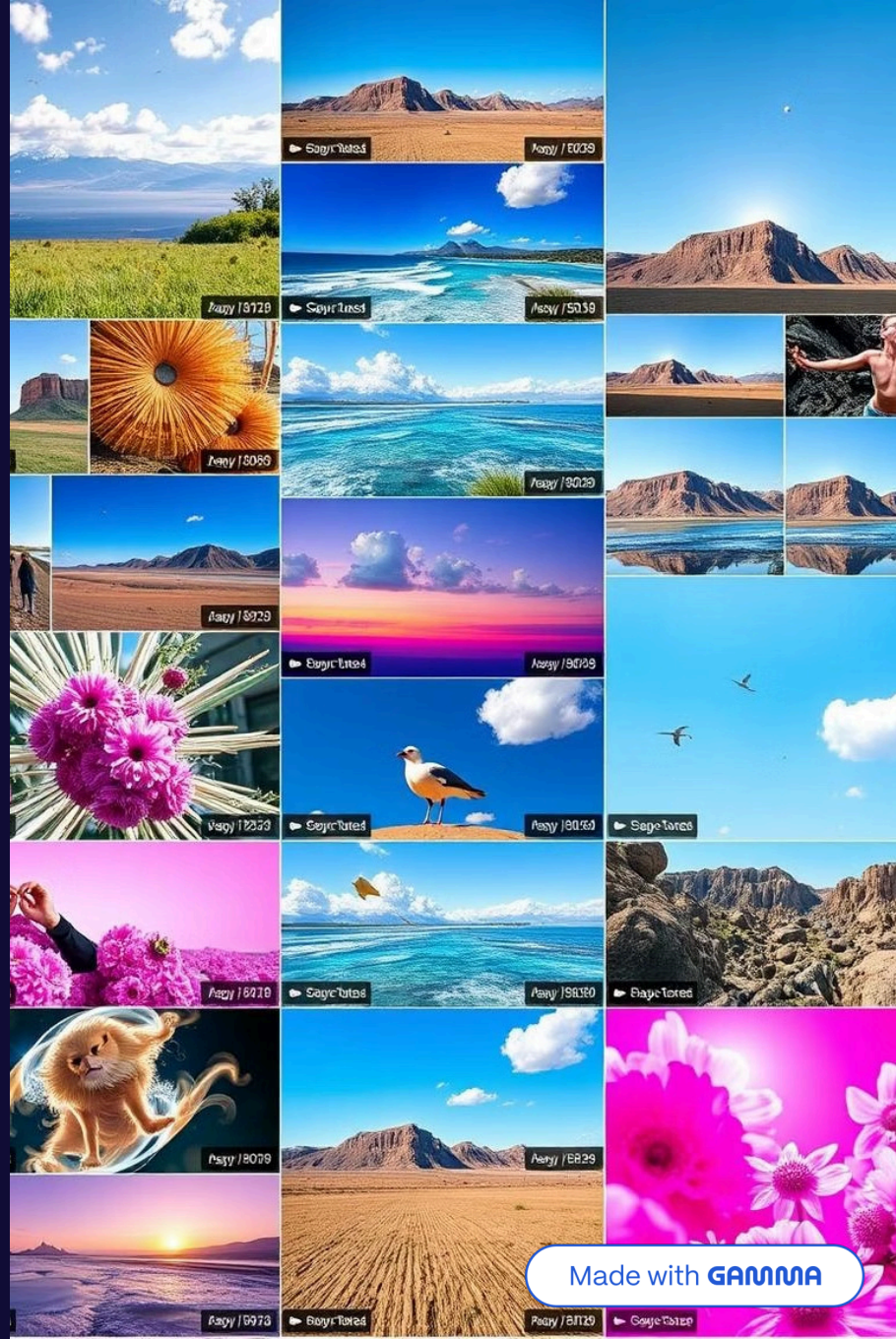
- Images from MS COCO, Flickr8k, and Flickr30k
- Videos from MSR-VTT and YouCook2

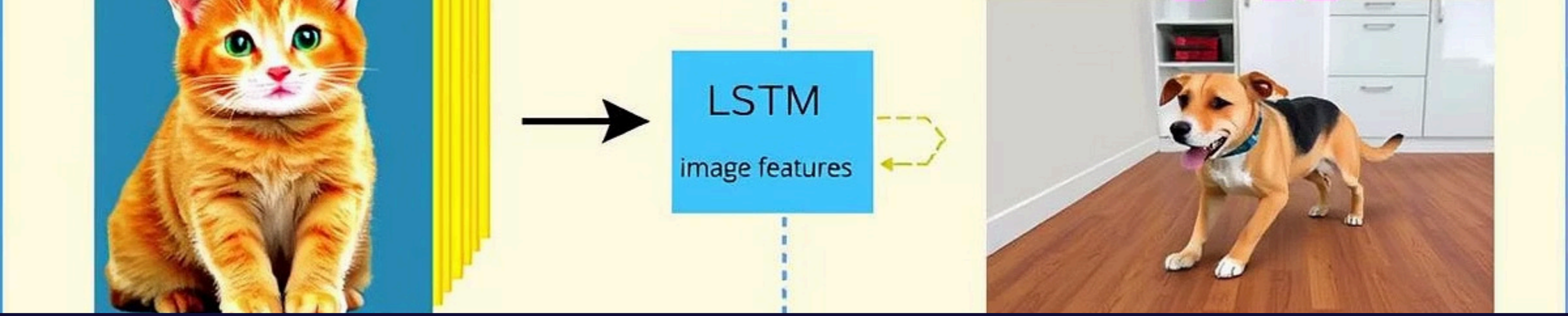
Annotation Process

Human experts provided descriptive captions aligned with visual content, ensuring high-quality ground truth for supervised learning.

Preprocessing

- Resizing images and extracting frames from videos
- Text captions tokenized and cleaned
- Optional feature extraction for enhanced consistency





Model Architecture

For Images

Encoder: A Convolutional Neural Network (CNN) such as ResNet or EfficientNet extracts rich feature representations from the input image.

Decoder: A Recurrent Neural Network (RNN) like LSTM or GRU, or a Transformer, generates a caption word-by-word based on the encoded image features.

Attention Mechanism: Helps the decoder focus on different parts of the image while generating each word.

For Videos

Frame Feature Extraction: Use a 3D CNN (e.g., C3D, I3D) or a pre-trained 2D CNN applied to selected frames to extract temporal-spatial features.

Sequence Modeling: Feed the sequence of frame features into an LSTM or Transformer to generate coherent video captions.

Temporal Attention: Dynamically focuses on relevant frames at each step while generating the caption.

Both image and video captioning models follow an Encoder-Decoder architecture with enhancements like Attention for better context understanding.

Feature Extraction with Convolutional Neural Networks

Spatial Feature Encoding

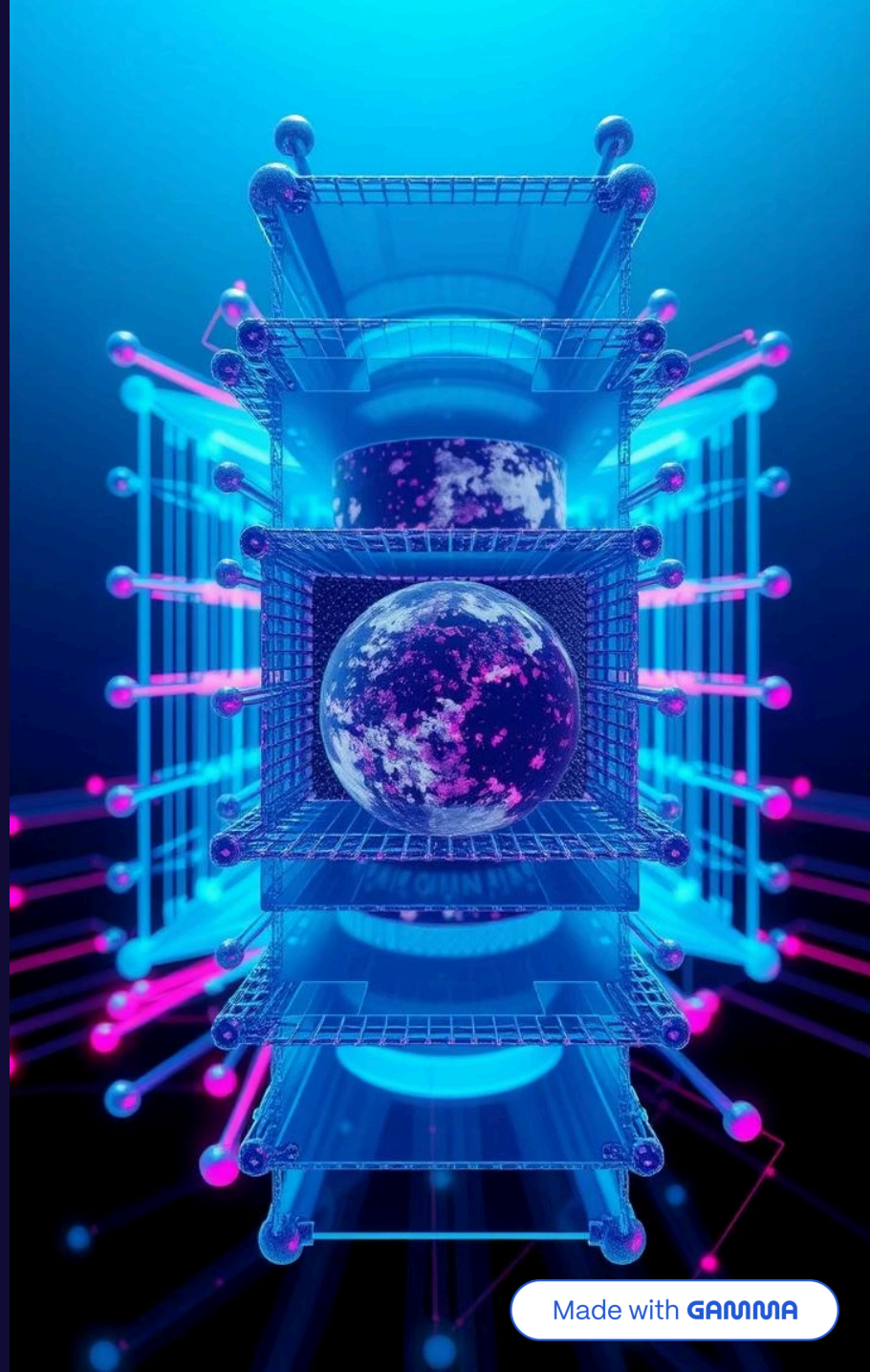
CNNs encode images by extracting hierarchical spatial features from low-level edges to high-level objects and scenes.

Video Frame Representation

Sequential frames are processed by CNNs independently to capture visual information before temporal modeling.

Pretrained Networks

Models like ResNet and Inception, pretrained on large image repositories, provide rich feature embeddings enhancing caption quality.



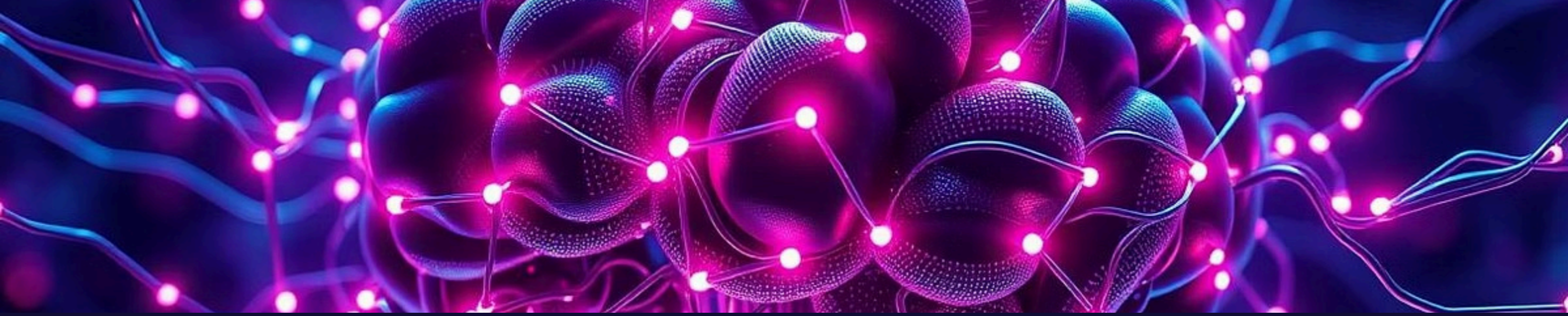
Sequence Modeling Using LSTMs

Handling Temporal Dependencies

LSTMs capture temporal context in sequences, maintaining memory of previous content to influence current word generation.

Generating Captions

The model predicts the next word token based on extracted visual features and previous tokens, producing coherent sentences.



Transformer-Based Caption Generation

1

Multi-Head Attention

Allows the model to focus on different image regions or temporal moments simultaneously for contextual understanding.

2

Positional Encoding

Integrates sequence order information crucial for accurate language modeling over video frames.

3

Stacked Layers

Enables deep feature interactions between visual inputs and linguistic outputs, improving caption coherence.

Evaluation Metrics

METEOR

Assesses harmonic mean of precision and recall, incorporating synonyms.

BLEU

Measures n-gram precision against reference captions.

CIDEr

Focuses on consensus in consensus captions by weighting relevant n-grams with TF-IDF.

Key Findings: Transformer Model Improvements

30%

Caption Accuracy Boost

Transformers showed a 30% improvement in descriptive accuracy over LSTM baselines.

40%

Context Understanding

Video captioning models better captured long-range temporal context, improving coherence by 40%.

25%

Preserving Dependencies

Spatial and temporal dependencies were maintained more effectively, enhancing caption relevance by 25%.

Conclusions and Next Steps

1 **Leverage Transformers More**

Further optimization and larger datasets could push caption quality closer to human levels.

2 **Integrate Multimodal Context**

Incorporating audio and scene metadata can enrich video captioning contextuality.

3 **Real-Time Applications**

Focus on efficiency for deployment in real-time visual content description systems.