Comparing Image Captioning Models: VGG16 and DenseNet 201

Srilekha Rayedi and Harshith Makkapati, *University of Houston*

Motivation Applications

Creating captions for images involves computer vision and natural language processing to form textual interpretations of visual content. In this project we are evaluating two models(VGG16 and DenseNetx 201) for an image captioning system. It is a Combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) to process and narrate visual data. We aim to distinguish how different image captioning is by comparing it with two models and utilizing the Flickr 8k dataset for validation. Our approach involves using a pretrained CNN to identify crucial features in images and integrating these features into the LSTM part of our model, which is responsible for generating consistent and context-aware captions.

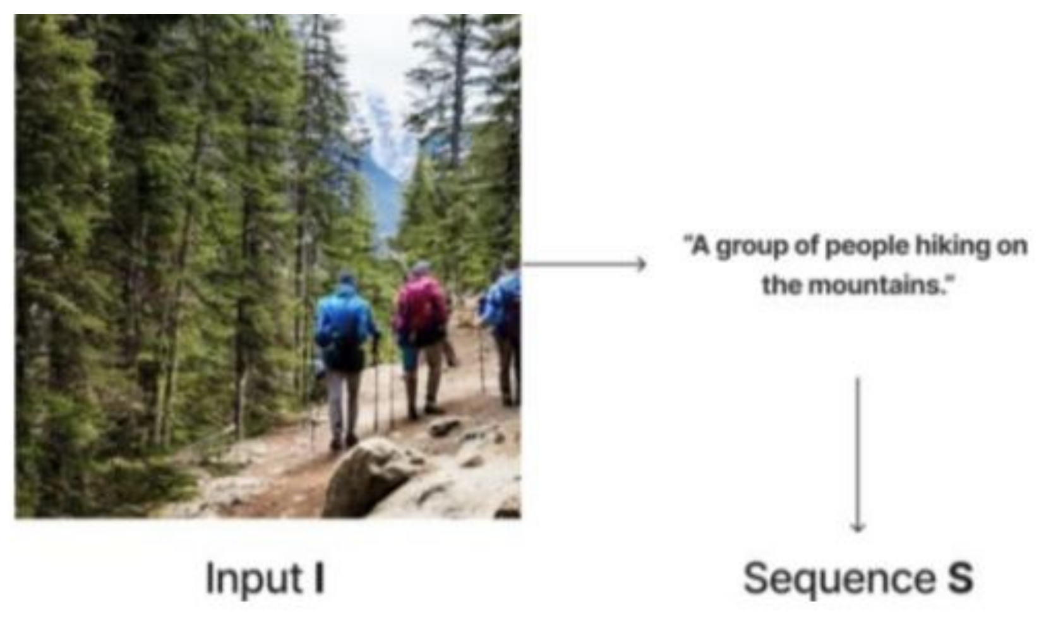


Figure 1: Model generating the caption for the image.

Formulation

Image Captioning with Deep Learning Architectures

Objective

- Develop an image captioning system using **CNNs** (e.g., VGG16 and DenseNet201) for feature extraction and **LSTM networks** for sequence generation.
- Minimize the overall loss while ensuring accurate caption generation and maintaining computational efficiency.

Goal: Minimize the loss over N time steps, where:

$$\min_{x} \sum_{t=1}^{N} Loss(x_t , y_t)$$

x_t: Model predictions at time t
y_t: Ground truth word at time t

The loss function is typically cross-entropy loss, which measures the difference between the predicted word probabilities and the actual word.

$$F = CNN(Image) + LSTM(Text\ Embedding)$$

Feature Fusion:

Visual Features: CNN(Image): Extracted from the image using a CNN model like VGG16 or DenseNet201.

Textual Context: LSTM(Text Embedding)): Generated by embedding words and passing them through an LSTM to capture sequential dependencies.

The fusion ensures that both image content and textual context contribute to generating the next word in the caption.

Input Processing:. Resize input images to 224 × 224 224×224.
Tokenize and pad captions to a fixed length.

Feature Extraction: Use pre-trained CNNs (e.g., VGG16, DenseNet201) to extract visual features.

Caption Generation: An LSTM decoder generates captions sequentially by combining CNN-extracted image features with word embeddings at each step.

Output: A complete caption generated word by word.

Existing Methods

Early Image Annotation Techniques [Before 2010]

- Manual extraction of visual elements (shapes, colors) combined with template-based systems.
- **Challenges**: Limited contextual understanding and generalization, leading to simplistic captions.

Neural Image Captioning (NIC) [Vinyals et al., 2015]

- End-to-end trainable model using:
  - **CNNs** for feature extraction.
  - **LSTMs** for generating descriptive captions.
- **Advantages**: Human-like captions with improved accuracy using large datasets like MS-COCO.

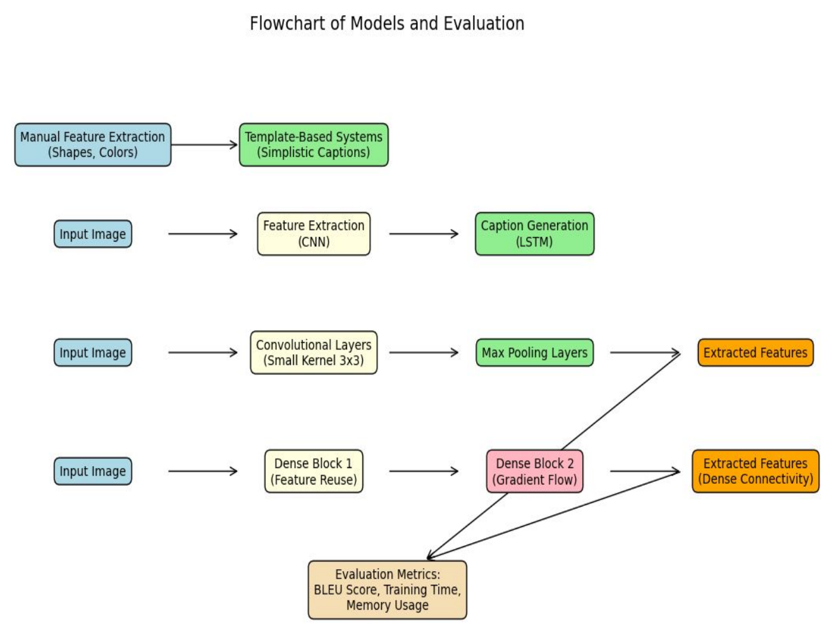
VGG16 [Simonyan and Zisserman, 2014]

- Deep CNN with sequential layers, small kernel sizes (3x3), and max pooling.
- **Strength**: Effective feature extraction for complex image tasks.
- **Limitations**: Higher computational cost compared to simpler models.

DenseNet201 [Huang et al., 2017]

- Introduces dense connections between layers for:
  - Better gradient flow.
  - Enhanced feature reuse.
- **Strength**: Improved performance for complex image features.
- **Limitations**: Increased training time due to dense connectivity.

Flowchart of Models and Evaluation


Manual Feature Extraction (Shapes, Colors)
Template-Based Systems (Simplistic Captions)
Input Image
Feature Extraction (CNN)
Caption Generation (LSTM)
Input Image
Convolutional Layers (Small Kernel 3x3)
Max Pooling Layers
Extracted Features
Input Image
Dense Block 1 (Feature Reuse)
Dense Block 2 (Gradient Flow)
Extracted Features (Dense Connectivity)
Evaluation Metrics: BLEU Score, Training Time, Memory Usage

Methods	Strengths	Weaknesses
Early Methods	Simplicity	Lacked context and generalization
NIC	Human-like captions, end-to-end design	Requires large datasets
VGG16	Efficient feature extraction	High computational cost

Algorithm Design

Framework and Architecture for CNN-LSTM Image Captioning

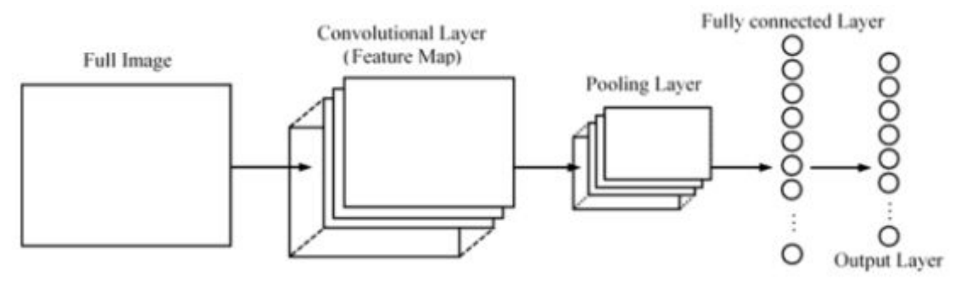


Figure 2: Basic structure of Convolutional Neural Network(CNN)

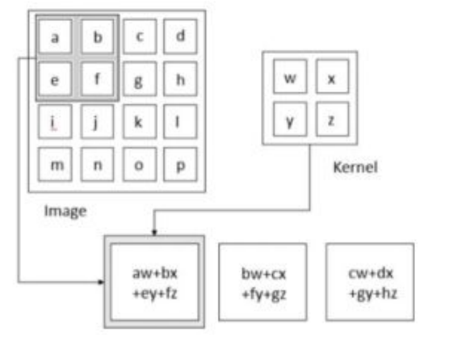


Figure 3: Convolution Layer Operation

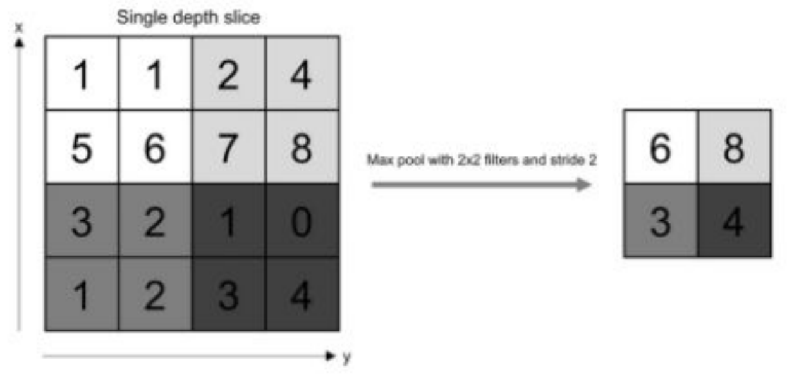


Figure 4: Max Pooling Operation

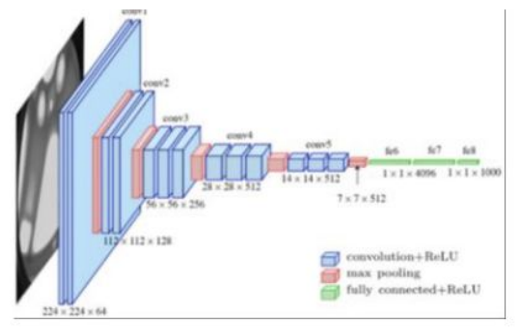


Figure 5: VGG16 architecture

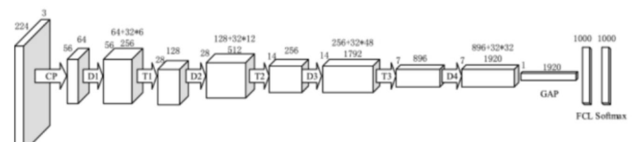


Figure 6: DenseNet201 architecture

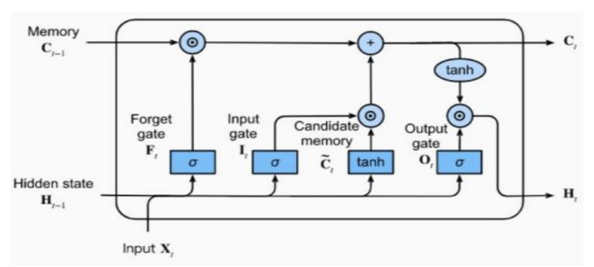


Figure 7: Architecture of LSTM Unit

Benefits:

- **Accuracy**: High BLEU score with coherent captions.
- **Scalability**: Adaptable to different datasets.
- **Interpretability**: Clear role of CNN for features and LSTM for sequences.

Challenges:

- **Feature Representation**: Balancing compactness with detail in CNN features.
- **Sequence Modeling**: Handling long-term dependencies in captions.
- **Computational Efficiency**: High memory usage of DenseNet201.

Performance Analysis

Loss Score of VGG16 and DenseNet 201

Metric	VGG16	DenseNet201
Initial Loss	5.7931	5.6745
Final Loss	2.1720	3.4736
Observation(Loss)	Significant reduction in loss; better generalization and learning.	Slower learning and higher final loss, indicating slower adaptation.
Generated Captions	More accurate and contextually relevant captions.	Less accurate captions, occasionally struggling with coherence.
Training Efficiency	Required less computational time due to simpler architecture.	Longer training time due to dense connections but better gradient flow.
Memory Usage	Lower memory usage.	Higher memory consumption due to feature reuse.
BELU Score	~65%	~60%

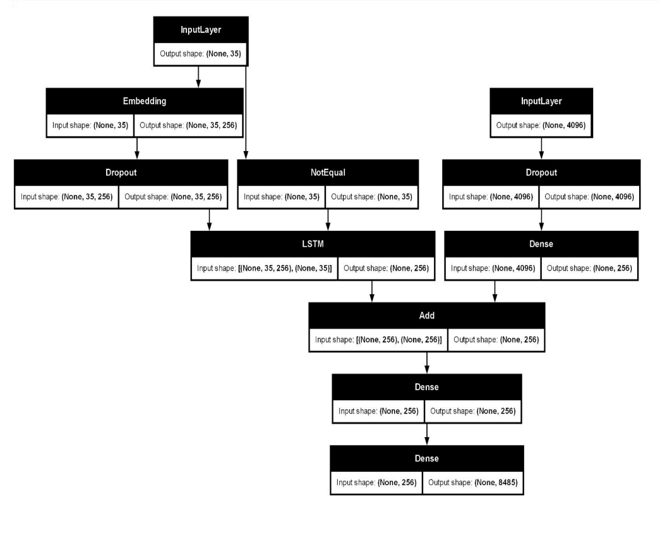


Figure 8: The implementation of VGG16 model

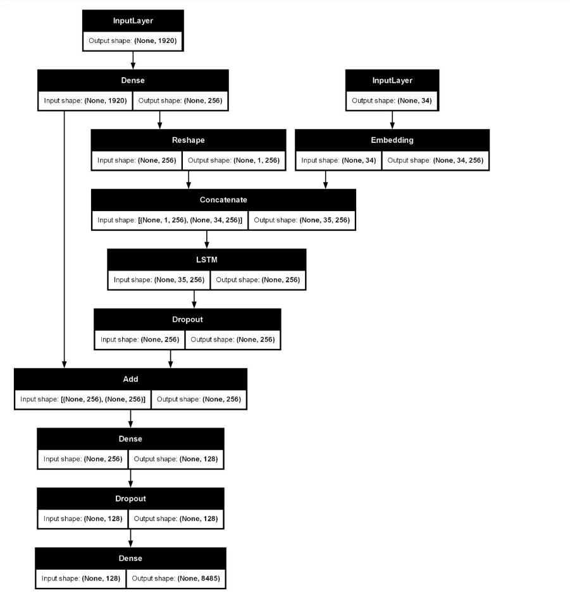
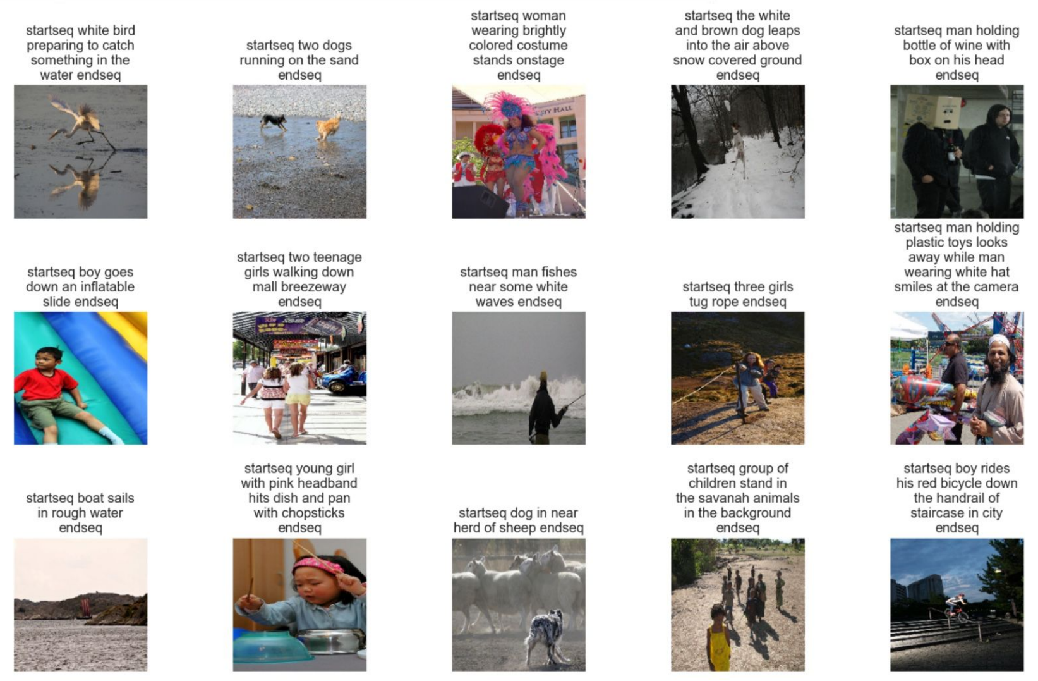


Figure 9: The implementation of DenseNet201 model


startseq white bird preparing to catch something in the water endseq
startseq two dogs running on the sand endseq
startseq woman wearing brightly colored costume stands onstage endseq
startseq the white and brown dog leaps into the air above snow covered ground endseq
startseq man holding bottle of wine with box on his head endseq
startseq boy goes down an inflatable slide endseq
startseq two teenage girls walking down mall breezeway endseq
startseq man fishes near some white waves endseq
startseq three girls tug rope endseq
startseq man holding plastic toys looks away while man wearing white hat smiles at the camera endseq
startseq boat sails in rough water endseq
startseq young girl with pink headband hits dish and pan with chopsticks endseq
startseq dog in near herd of sheep endseq
startseq group of children stand in the savanah animals in the background endseq
startseq boy rides his red bicycle down the handrail of staircase in city endseq

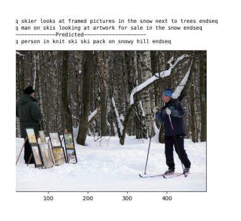
Experiments

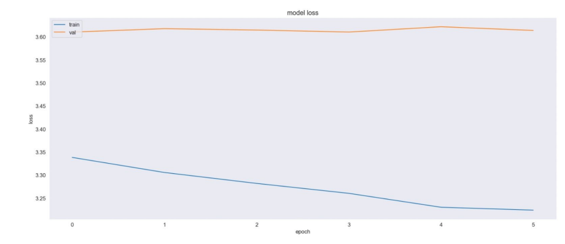
Dataset

- **Flickr8k**: 8,000 images with human-generated captions.
- **Split**: 6,000 (training), 1,000 (validation), 1,000 (testing).
- **Preprocessing**: Resize to 224x224, clean captions, and add `<start>/<end>` tokens.

Evaluation

- **Metrics**:
  - BLEU Scores: Assess caption quality.
  - Training Time & Memory Usage: Measure computational efficiency.


skier looks at framed pictures in the snow next to trees endseq
man on skis looking at artwork for sale in the snow endseq
Predicted
person in knit ski ski pack on snowy hill endseq



Model Training
Feature Extraction:

- Models: VGG16 and DenseNet201 (pre-trained on ImageNet).
- Extract features from the last convolutional layer.
- **Caption Generation**:
  - LSTM decodes features to generate captions.
- **Training Details**:
  - **Batch Size**: VGG16 (32), DenseNet201 (64).
  - **Optimizer**: Adam with learning rate decay.
  - **Loss Function**:
- **Environment**: Jupyter Notebook.
- **Resources**: CPU and GPU for training.

This study systematically evaluated the capabilities of sophisticated deep learning architectures, notably VGG16 and DenseNet201, in the field of picture captioning. We proved, using the Flickr8k dataset, which combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks to compare the relevancy of generated picture captions. The VGG16 model, with its deep yet simple architecture, has demonstrated impressive ability to extract precise visual features, whereas DenseNet201's complicated connection pattern has proven critical in guaranteeing robust feature propagation and successful learning. Our findings confirm that the strategic combination of these models enables a more sophisticated comprehension and development of natural language descriptions of pictures, therefore bridging a significant gap between visual data perception and language representation. This innovation not only advances academic research in picture captioning, but it also prepares the way for practical applications in a variety of disciplines, including automated surveillance, assistive devices for the visually impaired, and improved engagement with digital media.

References

[1] 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 (pp. 3156-3164).

[2] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[3] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. (2018). A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 51, 6, Article 118 (February 2019), 36 pages. https://doi.org/10.1145/3295748

[4] Md. S. Takkar, A. Jain and P. Adlakha, "Comparative Study of Different Image Captioning Models," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1366- 1371, doi: 10.1109/ICCMC51019.2021.9418451.

[5] Zhiyong Cui, Ruimin Ke, Ziyuan Pu, Yinhai Wang,Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values,Transportation Research Part C: Emerging Technologies,Volume 118,2020,1026