Image Caption Generation Report

Deep Learning Assignment 3

1. Introduction

The goal of this project is to develop an image captioning system that combines computer vision (CNN) and natural language processing (LSTM) techniques. The model learns to generate human-readable descriptions for input images using the Flickr8K dataset containing 8,000 images with 5 captions each.

2. Dataset

- **Dataset**: Flickr8K
- **Training Images**: 7,000
- **Validation Images**: 1,000
- **Test Images**: 200
- **Caption Statistics**:
  - Average caption length: 12-25 words
  - Vocabulary size: 8,765 unique words
  - Maximum sequence length: 35 tokens

3. Methodology

3.1 Model Architecture

Components:

1. **Image Encoder**: ResNet-50 (pretrained on ImageNet)
2. **Text Decoder**: 2-Layer LSTM with Attention
3. **Combination**: Feature fusion via concatenation

3.2 Preprocessing

Image Processing:

- Resize to 224×224
- Normalization: $\mu=[0.485,0.456,0.406]$, $\sigma=[0.229,0.224,0.225]$

Text Processing:

- Lowercase conversion
- Added <sos> and <eos> tokens
- Removed special characters
- Word tokenization

4. Training Details

Parameter	Value
Epochs	100
Batch Size	128
Learning Rate	5e-4
Optimizer	AdamW
Loss Function	CrossEntropy
Training Time	4.2 hrs

Key Techniques:

- Mixed Precision Training
- Learning Rate Scheduling
- Early Stopping (Patience=5)
- Gradient Scaling

5. Results

5.1 Performance Metrics

Metric	Training	Validation
Loss	1.24	2.87
BLEU-1	0.62	0.54
BLEU-4	0.31	0.23

5.2 Sample Predictions

Example 1:

- **Image**: Beach scene
- **Actual**: "A group of people playing volleyball on sandy beach"
- **Predicted**: "People are playing game on beach with net"

Example 2:

- **Image**: City street
- **Actual**: "Busy city street with yellow taxis and pedestrians"
- **Predicted**: "Urban road with cars and people walking"

6. Hyperparameter Analysis

Learning Rate Comparison

LR	Val Loss	Convergence Epochs
1e-3	3.45	28
5e-4	2.87	35
1e-4	3.12	52

Batch Size Impact

7. Challenges & Solutions

1. **Challenge**: Overfitting with small dataset
   Solution: Added dropout (0.5) and early stopping

2. **Challenge**: Long training time
   Solution: Implemented mixed precision training (40% speedup)

3. **Challenge**: Rare word handling
   Solution: Added <unk> token for words <5 frequency

8. Conclusion

- Achieved 54% BLEU-1 score on validation set
- Model successfully learns image-text relationships
- Attention mechanism helps focus on relevant image regions

Future Improvements:

- Use larger dataset (Flickr30K)
- Implement Transformer architecture
- Add beam search decoding