Vision to Text: An Image Captioning System

Curtis Neiderer, Divya Maheshkumar, Desiree Reid, Minal Ahir, Arundhati Ubhad

1. Final Topic Area

This project falls within the Natural Language Processing (NLP) domain, as it focuses on generating coherent, semantically accurate natural language captions. However, it also draws meaningfully from Computer Vision (CV) by requiring visual feature extraction from input images. We selected this topic because of the group's common interest in NLP-based generative models and our desire to apply key techniques from both areas covered in this course. This multimodal task allows us to engage with foundational generative AI concepts, such as encoder-decoder architectures, transfer learning with pretrained models (e.g., ResNet-50), and latent space analysis to evaluate semantic fidelity. It also supports deeper exploration of key course outcomes, including the ability to design and optimize generative architectures, evaluate their performance across domains using both traditional and embedding-based metrics, and analyze how generative systems learn and generalize through feedback.

2. Dataset Description

The following two datasets have been identified to support the development of the project.

- **Flickr8k** (https://www.kaggle.com/datasets/adityajn105/flickr8k)
  The Flickr8k dataset is a collection of 8,092 images with 5 captions per image describing the salient entities and events. It is widely used in computer vision and natural language processing research, particularly for tasks like image captioning. The dataset was collected from user-uploaded Flickr photos and was initially released by the University of Illinois in 2013.
- **Flickr30k** (https://www.kaggle.com/datasets/awsaf49/flickr30k-dataset)
  The Flickr30k dataset is a collection of 31,783 images with 5 captions per image describing the salient entities and features. It is an expansion of the original Flickr8k dataset and is a widely used benchmark in the field of computer vision, specifically for sentence-based image description and visual-linguistic tasks. The dataset was collected from user-uploaded Flickr photos and was initially released by the University of Illinois in 2014.

3. Model Selection

The image captioning system will be developed using a deep learning-based encoder-decoder architecture that translates image features into coherent natural language descriptions.

- **Image Feature Extraction (Encoder)**
  We will use a pre-trained ResNet-50 Convolutional Neural Network (CNN) as the encoder to extract high-level visual features from input images. The pre-trained ResNet-50 CNN

model captures rich spatial and semantic information due to its expansive training on large datasets like ImageNet.

- **Caption Generation (Decoder)**
  For the decoder, we will employ an LSTM (Long Short-Term Memory) model to generate a sequence of words (caption) based on embedded image features. At each step, the LSTM will predict the next word in the caption, conditioned on the previously generated words and the visual context.

4. Research Questions

This project aims to investigate the following research questions.

1. How accurately do image captioning models preserve the semantic content of an input image?
2. Where do modern generative captioning models hallucinate or misinterpret visual content?
3. Can embedding-based metrics or visual attention maps help quantify caption fidelity beyond BLEU and METEOR scores?
4. What failure modes occur when captioning under visual ambiguity (e.g., occlusions, cluttered scenes, etc.)?

5. Plan of Action

Data Preprocessing

The image preprocessing pipeline will consist of the following steps.

1. **Image Resizing**: Resizing all images to a fixed size of 224x224 pixels to match model input requirements as well as to standardize input and reduce computation load.
2. **Pixel Normalization**: Scale input images to the [0,1] range by dividing pixel values by 255, and then normalize it using the ImageNet channel-wise mean and standard deviation (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to ensure compatibility with the pretrained ResNet-50 model.
3. **(Optional) Data Augmentation (Training Only):** Helps the model generalize better
   - Center Cropping / Random Cropping
   - Random Flip (horizontal or vertical)
   - Random Rotation
   - Color Jitter

The caption preprocessing pipeline will consist of the following steps.

1. **Tokenize Captions**: Use sub-word tokenization to reduce out-of-vocabulary issues with rare words.
2. **Build Vocabulary**: Threshold low frequency words to lower memory usage and speed-up training and inference.

3. **Caption Length Truncation/Padding**: Truncate or pad captions to uniform length expected by model.

Model Implementation and Training

The feature extraction model (encoder) will be ResNet-50 CNN model with ImageNet weights and the caption generation model (decoder) will be an LSTM language model with the following architecture:

- Embedding Input Layer
- LSTM Decoder
- Dense Output Layer

The model training setup will be as follows:

- Loss Function: Categorical Cross-Entropy
- Optimizer: Adam (initial learning rate ~ 0.001)
- Batch Size: 64
- Epochs: ~20-50 with early stopping

Note: All training parameters are tunable, allowing these values to be adjusted as necessary to improve performance.

Experiment Design and Performance Evaluation

The following experiments have been compiled to address the research questions described above.

1. **Baseline Training and Evaluation (Supports Research Question 1)**
   Train CNN-LSTM model on Flickr8k and evaluate using automatic metrics (BLEU, METEOR, etc.)
2. **Error Analysis (Supports Research Questions 2 and 4)**
   Qualitative analysis of bad predictions, noting confusing categories (i.e., missing objects, hallucinations, grammatical errors).
3. **Semantic Fidelity Comparison (Supports Research Question 3)**
   Compute the correlation between human judgement and traditional metrics (e.g. BLEU and METEOR,etc.) compare it to the correlation between human judgement and embedding-based metrics (e.g., BERTScore, CLIPScore, etc.). Optional: If time permits, compute correlation between human judgement and the visual attention map alignment, then compare against traditional and embedding-based metrics.
4. **Generalization (Supports Research Question 1)**
   Assess the model's ability to generalize by evaluating its captioning performance on an out-of-domain dataset (e.g., Flickr30k) without additional fine-tuning. Performance will be measured using standard metrics (BLEU, METEOR, etc.) to determine how well the model transfers to visually and semantically diverse content beyond the training distribution.

Project Timeline

The approximate project schedule is as follows:

- Week 3: Initial Project Idea and Proposal (Milestone 0)
- Week 4-6: Project Idea Refinement
- Week 7-8: Dataset Selection; Model Selection; Metric Selection
- Week 8-9: Final Project Proposal (Milestone 1)
- Week 9-10: Data Preprocessing Pipeline (Milestone 2)
- Week 10-11: Model Implementation and Training Pipeline (Milestone 3)
- Week 11-12: Model Evaluation Pipeline and Performance Interpretation
- Week 12-14: Final Project Report Submission and Presentation (Milestone 4)

6. Team Contribution

The roles within the group align to the main project areas:

- **Data Engineer (Curtis):** Identifies relevant datasets, obtains the data and develops the preprocessing pipeline
- **Computer Vision Engineer (Divya):** Identifies model framework, implements and trains feature extraction model
- **NLP Engineer (Desiree):** Identifies model framework, implements and trains caption generation model
- **Performance Evaluation Engineer (Minal):** Identifies relevant performance metrics and develops evaluation pipeline
- **Visualization Engineer (Arundhati)**: Generates visualizations showcasing model performance in aggregate as well as select examples
- **All Members:** Literature Review, Collaborative Analysis and Report Writing

Note: These roles are flexible, and members may take on different responsibilities as needed to support the project's progress and foster collaboration within the group.