Vision to Text: An Image Captioning System

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

Project Description

This project aims to investigate the performance and limitations of deep generative models in producing accurate natural language captions for images. By leveraging a multimodal architecture combining computer vision (CV) and natural language processing (NLP), we will evaluate whether generated captions preserve the semantic and visual integrity of source images. Key focuses include defining a metric for "caption fidelity," analyzing training convergence, and examining how input variations impact caption generation and model robustness. The entire workflow will be developed and evaluated in a Jupyter Notebook, resulting in a streamlined, repeatable pipeline.

Problem Statement

The exponential growth of visual data has increased demand for automatic, meaningful image descriptions. Applications range from aiding visually impaired users and enhancing image search to improving e-commerce product captions, medical image reporting, and social media content tagging. Despite advances, generating context-aware captions remains challenging due to the semantic gap between visual content and natural language. Our project seeks to close this gap by building a captioning system that can understand the content of an image and describe it accurately in natural language.

Dataset Selection

The following datasets will be used for training and evaluation.

1. MS COCO 2014/17 (https://paperswithcode.com/dataset/coco):
    - Widely used benchmark for image captioning.
    - Consists of 328K images.
2. Flickr8k / Flickr30k (https://paperswithcode.com/dataset/flickr-8k; https://paperswithcode.com/dataset/flickr30k):
    - Smaller datasets suitable for quick prototyping.
    - Ideal for early model testing before scaling up to MS COCO.

All datasets will be resized and preprocessed for consistency and manageable runtime.

Background

Image captioning combines vision and language modeling and has been widely studied in recent years. Traditional approaches used template-based methods, but the field shifted significantly with the introduction of deep learning.

A foundational model, "Show and Tell" by Vinyals et al. (2015), introduced an encoder-decoder architecture using CNNs to process images and RNNs to generate captions. This was extended by Xu et al. (2015) with "Show, Attend and Tell", which added attention mechanisms to improve relevance and interpretability by focusing on image regions during captioning.

Recent progress includes BLIP (Li et al., 2022), which unified vision-language pretraining to support both understanding and generation tasks. BLIP-2 (Li et al., 2023) advanced this further by using a frozen image encoder and a large language model decoder, improving modularity and performance on multimodal tasks.

References

1. Vinyals et al. (2015). *Show and Tell*. https://arxiv.org/abs/1411.4555
2. Xu et al. (2015). *Show, Attend and Tell*. https://arxiv.org/abs/1502.03044
3. Li et al. (2022). *BLIP*. https://arxiv.org/abs/2201.12086
4. Li et al. (2023). *BLIP-2*. https://arxiv.org/abs/2301.12597
5. Liu et al. (2022). *A Frustratingly Simple Approach*. https://arxiv.org/abs/2201.12723

Methodology

The image captioning system will be developed using a deep learning-based encoder-decoder architecture that translates image features into coherent natural language descriptions. The pipeline consists of four main components: image feature extraction, sequence generation, model training and evaluation.

1. **Image Feature Extraction (Encoder)**
   We will use a pre-trained Convolutional Neural Network (CNN), such as ResNet-50 or InceptionV3, as the encoder to extract high-level visual features from input images. These CNN models, pre-trained on large datasets like ImageNet, are capable of capturing rich spatial and semantic information. Alternatively, we will explore using pretrained multimodal models such as BLIP or BLIP-2 to leverage their powerful vision-language representations.

2. **Caption Generation (Decoder)**
   For the decoder, we will employ a Recurrent Neural Network (RNN), specifically an LSTM (Long Short-Term Memory) model, to generate a sequence of words (caption) based on the image features. We may also experiment with Transformer-based decoders, which have shown improved performance in language generation tasks. The LSTM decoder will be initialized with the encoded image features either by feeding them as the first input token, using them to initialize the hidden state, or via an attention mechanism that allows the decoder to dynamically focus on different parts of the image during generation. At each time step, the LSTM will predict the next word in the caption, conditioned on the previously generated words and the visual context.

3. **Training Approach**
   We will train the model using supervised learning on pairs of images and ground truth captions. Pretrained models will be fine-tuned for improved performance and faster convergence. Training the captioning model involves learning to align image features with natural language in a supervised manner. We will use **paired** image-caption datasets, such as MS COCO, Flickr30k where each image is annotated with one or more humanized descriptions. To improve robustness, we will also consider data augmentation techniques such as random cropping or rotation.

4. **Evaluation**
   To evaluate the quality of the generated captions, we will use established NLP metrics (BLEU, METEOR, etc.) as well as qualitative analysis through visual inspection of image-caption pairs. These metrics will be computed on widely used benchmark datasets such as MS COCO and Flickr30k, enabling comparisons within the datasets. This dual approach ensures that the system not only performs well numerically but also generates captions that are meaningful and humanizing.

Expected Outcome

Our project aims to contribute a clearer theoretical and empirical understanding of semantic fidelity in multimodal generative models, particularly within image captioning tasks. The primary outcomes we anticipate include:

1. **A structured framework** for evaluating whether generated captions accurately reflect the sematic content of input images, beyond traditional n-gram or surface level metrics.

2. **Empirical analysis** of where and how current models (e.g., BLIP, CLIP-based generators) diverge from visual input, including insights into latent space misalignment and semantic hallucination.

3. **Recommendations for improved evaluation techniques,** potentially involving embedding-based or image-aware fidelity metrics.

4. **A small curated diagnostic benchmark,** featuring edge cases (e.g., object omission, attribute errors) to support future captioning evaluation research.

While our project is theoretical in nature, these outcomes hold practical value in several real-world domains where the accuracy of AI-generated captions is critical. These include:

- **Acessibilitily Tools**: Visually impaired users depend on accurate image captions for screen reader navigation. Incorrect or hallucinated captions can be misleading or harmful.
- **E-commerce**: product images require precise, descriptive captions for effective search, categorization, and recommendation. Errors in captioning can directly affect sales and user trust.
- **Content Moderation:** Social media platforms rely on captions to flag or review visual content. Misaligned captions can lead to false positives or missed violations.
- **Digitial Media Indexing**: Platforms such as Pinterest or Instagram use generated tags and captions for visual context indexing, affecting discoverability and personalization algorithms.
- **Medical Imaging (future work):** In healthcare settings, where captioning is being explored for radiology reports, high-fidelity generation is essential to prevent clinical errors

Team Roles

There are four main project roles, each contributing to different aspects of the pipeline.

1. **Data Engineer:** Leads Data Preprocessing and Management
2. **Computer Vision Engineer:** Leads Encoder Development
3. **NLP Engineer:** Leads Decoder Development
4. **Evaluation Engineer:** Leads Performance Evaluation and Visualization
   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.