# Project Report: Image Captioning on Flickr8k


- Sonu Soni
- Raghu Vamsi Sangala
- Arshad Jafri


## Background

In recent years, the field of artificial intelligence has witnessed significant progress in bridging the gap between computer vision and natural language processing. One of the key challenges at this intersection is image captioning, which refers to automatically generating natural language descriptions of images.

Traditionally, image captioning systems use a two-stage deep learning architecture:
1. A Convolutional Neural Network (CNN) as an encoder to extract visual features from an image.
2. A Recurrent Neural Network (RNN), typically an LSTM (Long Short-Term Memory) network, as a decoder to generate a caption based on those features.

This encoder-decoder framework has been widely adopted and serves as the foundation for many neural image captioning models.

In this project, we initially implemented such a CNN+LSTM architecture but later explored leveraging pretrained vision-language models like BLIP, which integrate both visual and language understanding in a single transformer-based architecture.

Image captioning plays a crucial role in applications such as:
- Enhancing accessibility for visually impaired individuals.
- Improving image retrieval systems.
- Assisting content moderation and social media analysis.
- Enabling human-computer interaction in multi-modal settings.

This project seeks to evaluate and compare these approaches on the Flickr8k dataset.

## Project Motivation

The motivation for this project stems from the desire to explore how pretrained vision-language models can perform on smaller, specialized datasets without additional fine-tuning. While large-scale datasets like MSCOCO are commonly used for training, we chose the Flickr8k dataset due

to its manageable size and academic popularity.

Key motivations include:
- Understanding zero-shot capabilities of pretrained models.
- Measuring how well pretrained models generalize to new data.
- Exploring the effectiveness of BLEU metrics for caption evaluation.
- Gaining hands-on experience with Hugging Face Transformers and image captioning pipelines.

Ultimately, this project aims to highlight both the potential and limitations of using off-the-shelf models for image captioning on smaller datasets.

## Dataset

We used the Flickr8k dataset, a benchmark dataset for image captioning research. It contains:
- 8,000 images covering diverse real-world scenes.
- Each image is paired with 5 human-annotated captions describing the content.

Dataset statistics:
- Average caption length: ~10 words
- Vocabulary size: ~8,000 unique words

Dataset was downloaded and extracted using Python code, and captions were loaded into a Pandas DataFrame for analysis. Each row contains an image filename and a corresponding caption.

## Method

The project involved two main approaches for image captioning:

1. CNN + LSTM pipeline:
- We initially implemented a classical encoder-decoder model:
  - CNN Encoder: a pretrained CNN (e.g., ResNet or EfficientNet) was used to extract a feature vector from input images.
  - LSTM Decoder: an LSTM network was trained to generate captions based on the image feature embedding and ground-truth captions during training.

Training involved teacher forcing, cross-entropy loss, and gradual learning rate decay. However, this approach faced challenges in convergence and performance on a small dataset.

2. Pretrained BLIP model:
- To leverage state-of-the-art vision-language understanding, we utilized the BLIP (Bootstrapping Language-Image Pretraining) model via Hugging Face Transformers.
- BLIP integrates a Vision Transformer (ViT) encoder and a transformer-based language decoder,

pretrained on large-scale vision-language data.
- Using BLIP enabled zero-shot captioning on Flickr8k without additional training.

The implementation included loading the pretrained model, processing images, generating captions, and evaluating using BLEU metrics.

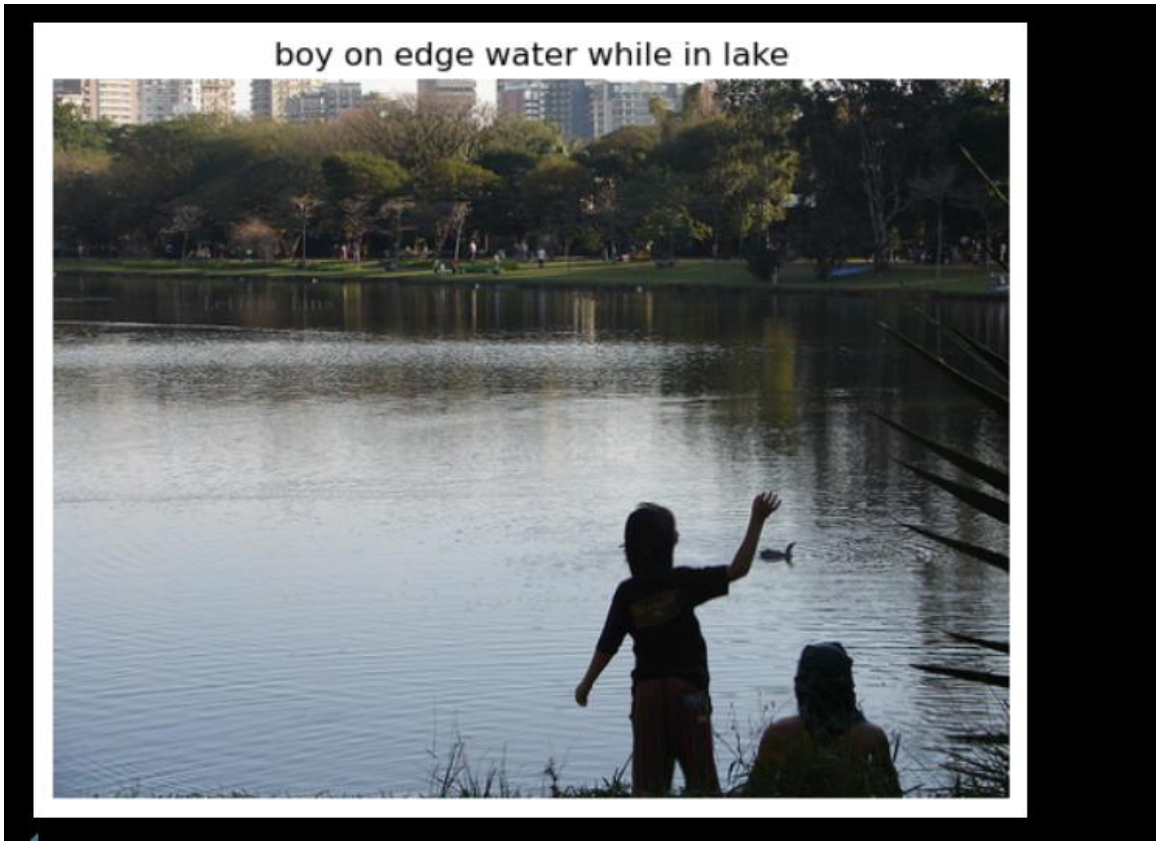
## Code Snippets

Example code for caption generation and BLEU evaluation was included in the Jupyter notebook. Key steps include loading the BLIP model, preprocessing images, generating captions, and evaluating with BLEU. See notebook for complete code.

## Experiments and Results

We ran caption generation on a subset of test images and computed BLEU scores against the 5 provided ground-truth captions per image.

Examples:



Results:

| Metric | Score |
|---|---|
| BLEU-1 | 0.2115 |
| BLEU-2 | 0.0656 |
| BLEU-3 | 0.0361 |
| BLEU-4 | 0.0232 |

Observations:
- BLEU-1 shows moderate unigram overlap.
- BLEU-4 is low, reflecting strict matching of longer n-grams.
- Generated captions were grammatically correct but often lacked specific details present in human annotations.

## Discussion

The project reveals important insights.

Initially, we implemented a CNN+LSTM encoder-decoder model from scratch, following the classical image captioning paradigm. While this approach provided valuable learning insights, it required extensive training to converge and was prone to overfitting due to the small size of Flickr8k.

By transitioning to a pretrained BLIP model, we significantly reduced training time and achieved more fluent captions out-of-the-box. However, BLEU scores remained modest due to differences in vocabulary and phrasing compared to Flickr8k's ground-truth captions.

Strengths:
- BLIP model produces fluent, grammatical captions without fine-tuning.
- Zero-shot capability generalizes reasonably well on Flickr8k.

Limitations:
- BLEU metric penalizes synonyms and paraphrases despite semantic correctness.
- Model struggles with fine-grained scene details (e.g., object counts, relationships).

Improvement avenues:
- Fine-tuning on Flickr8k captions.
- Incorporating image augmentation.
- Using alternative metrics like METEOR or CIDEr.

- Applying beam search or top-k sampling in decoding.

This project demonstrates that pretrained models can serve as strong baselines even for small datasets, but custom fine-tuning is critical for task-specific performance.

## References

1. Hugging Face Transformers documentation: https://huggingface.co/transformers/
2. BLIP model card: https://huggingface.co/Salesforce/blip-image-captioning-base
3. Flickr8k dataset: https://github.com/awsaf49/flickr-dataset
4. BLEU metric: Papineni et al., "BLEU: a method for automatic evaluation of machine translation" (2002)