Image Captioning using BLIP Model

Taehyeon Kim (2024149040)

1 Introduction

In this project, I build a small end-to-end AI pipeline for image captioning, the task of generating a short natural-language description given a single input image. Image captioning is an interesting example of multimodal AI because it has to understand both visual content and language at the same time. Even on a small scale, it clearly shows the difference between simple hand-crafted rules and modern pretrained models.

To keep the project computationally manageable, I use a subset of about 2500 image–caption pairs from the Flickr8k dataset (Hodosh et al. [1]), each image having multiple human-written captions. I compare two approaches:

- A rule-based baseline that generates captions using simple image statistics such as brightness, dominant color channel, and edge density.
- An AI pipeline based on a pretrained BLIP image-captioning model (Li et al. [2]), which encodes the image with a vision backbone and decodes a caption token by token.

Both methods are evaluated on a held-out test set using automatic metrics (BLEU-1 and BLEU-2) (Papineni et al. [3]) with multiple reference captions per image, as well as qualitative examples.

2 Task Definition

- **Task description:** Image captioning on Flickr8k (Hodosh et al. [1])
- **Motivation:** Image captioning is a canonical example of multimodal AI, where a system must jointly reason about visual information and natural language.
- **Input / Output:** The input to the model is a single RGB image and the output is the caption about the input image
- **Success criteria:**I consider the project successful if The BLIP-based AI pipeline achieves higher BLEU-1 and BLEU-2 scores than the rule-based baseline on the held-out test set, using the BLEU metric introduced by Papineni et al. [3] and evaluated with multiple reference captions per image.

3 Methods

This section includes both the naïve baseline and the improved AI pipeline.

3.1 Naïve Baseline

- **Method description:** The model computes simple global image statistics (brightness, dominant color channel, edge density) and maps them via a few if–else rules to one of several fixed caption templates.

- **Why naïve:** It does not learn from data or recognize specific objects or actions; it only uses hand-crafted rules and a tiny set of generic sentence templates.
- **Likely failure modes:** It will often produce vague captions that miss key objects or events, and it can misdescribe images when low-level cues (e.g., color or texture) do not match the true semantic content.

3.2 AI Pipeline

- **Models used:** I use the pretrained BLIP image-captioning model (Li et al. [2])`Salesforce/blip-image-captioning-base`, which combines a vision encoder with a text decoder to generate captions from images.
- **Pipeline stages:** The AI pipeline first preprocesses the input image, feeds it through the BLIP vision encoder, then autoregressively decodes a caption token by token using beam search, and finally evaluates the generated sentence with BLEU-1/2 against multiple reference captions.
- **Design choices and justification:** I chose BLIP because it is a widely used vision–language model that can be used off-the-shelf on a small dataset, giving strong captions with minimal training, and I kept the pipeline simple (no fine-tuning) to focus on comparing a strong pretrained model against a much weaker rule-based baseline under the same evaluation protocol.

4 Experiments

4.1 Datasets

- **Source:** A subset of the Flickr8k image–caption dataset introduced by Hodosh et al. [1] and accessed via the Hugging Face `jxie/flickr8k` dataset.
- **Total examples:** I use 2,500 images, each with up to five human-written reference captions.
- **Train/Test split:** The 2,500 images are randomly shuffled and split into 70% training, 15% validation, and 15% test using a fixed random seed for reproducibility.
- **Preprocessing steps:** Images are converted to RGB, resized to 256×256, and normalized; for the BLIP model I use its official processor for image and text preprocessing.

4.2 Metrics

For quantitative evaluation of caption quality, I use BLEU-1 and BLEU-2, which measure the overlap of 1-grams and 2-grams between the generated caption and human reference captions, following the BLEU metric introduced by Papineni et al. [3].

In my setup, each test image has up to five human-written reference captions, and BLEU is computed at the sentence level using all available references per image. BLEU-1 captures whether the model predicts roughly the right content words, while BLEU-2 is slightly stricter and also rewards short word sequences, making these metrics suitable for comparing the rule-based baseline and the BLIP-based model on this simple image-captioning task.

4.3 Results

Method	BELU-1	BELU-2
Baseline	0.40	0.11
AI Pipeline	0.56	0.40

Qualitative examples. Below are three representative examples from the test set.

- **Example 1**
  Ground-truth: "A girl with black gloves is running ."
  Baseline: "A simple scene with a few large regions."
  BLIP: "A woman wearing a purple shirt."

- **Example 2**
  Ground-truth: "Many people look over the side of a bridge ."
  Baseline: "A simple scene with a few large regions."
  BLIP: "A group of people standing on a bridge over a river"

- **Example 3**
  Ground-truth: "A brown dog digging a hole ."
  Baseline: "A busy scene with many edges and details."
  BLIP: "A dog digging a hole in the ground."

5 Reflection and Limitations

Overall, the BLIP-based pipeline worked better than I expected in terms of producing fluent, detailed captions, especially compared to the very simple rule-based baseline. However, I was surprised by how high the baseline BLEU-1 and BLEU-2 scores were relative to BLIP, given that the baseline captions looked much worse to me qualitatively. Implementing the pipeline also involved some practical difficulties, such as understanding the Flickr8k dataset format, handling multiple reference captions, and debugging batching and preprocessing for the pretrained model. Through the experiments, I realized that BLEU does not fully capture my intuitive notion of "quality": the metric can give a surprisingly decent score to generic captions that happen to share common unigrams or bigrams with the references, while penalizing more specific paraphrases that use different wording. In several examples, BLIP produced semantically accurate sentences that used synonyms or slightly different phrasing than the ground-truth, but BLEU still gave them relatively low scores. If I had more time or compute, I would like to try additional metrics that are better aligned with semantic similarity (e.g., CIDEr or BERT-based scores) and also include a small human evaluation. I would also explore fine-tuning BLIP on the training split or increasing the dataset size to see how much the captions and metrics can be improved beyond the off-the-shelf model.

References

[1] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013. URL https://jair.org/index.php/jair/article/view/10833.

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C H Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. Association for Computational Linguistics, 2002. URL https://aclanthology.org/P02-1040/.