BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

Yash Parmar

yqp5233@psu.edu

1. Introduction (FINAL – Template-Aligned, Short)

Vision–Language learning aims to develop models that jointly process visual content and natural language. A core task in this area is **image captioning**, which requires generating natural language descriptions that accurately reflect the content of an image. Image captioning is widely used in applications such as accessibility tools, image retrieval systems, and automated media annotation.

Earlier approaches to image captioning typically combined convolutional neural networks (CNNs) for visual feature extraction with recurrent neural networks (RNNs) for sequence generation. While effective, these methods were limited in their ability to capture complex visual–textual relationships. More recent Vision–Language Pretraining (VLP) models leverage large-scale image–text data and Transformer architectures to improve multimodal representations, but many focus primarily on understanding tasks rather than text generation.

The **BLIP (Bootstrapping Language-Image Pre-training)** framework addresses this limitation by proposing a unified architecture that supports both vision-language understanding and generation. In this project, we fine-tune a pretrained BLIP model for the **image captioning task** using the **Flickr8k dataset** and evaluate its performance using BLEU scores. We further analyse the effect of training duration and discuss potential improvements for small-scale datasets.

2. Background and Related Work

Image captioning is a fundamental task in vision–language research that requires generating natural language

descriptions for images. Over the years, several approaches have been proposed, evolving from traditional encoder–decoder models to large-scale pretrained vision–language systems.

Early image captioning approaches followed a **CNN–RNN framework**, where a convolutional neural network (CNN) extracted visual features and a recurrent neural network (RNN) generated captions sequentially. One of the earliest and most influential works, *Show and Tell* [1], demonstrated that image captioning could be formulated as a sequence prediction problem conditioned on image embeddings. Although effective, these models relied on global image features and often failed to capture fine-grained visual details.

To overcome this limitation, **attention-based approaches** were introduced. Models such as *Show, Attend and Tell* [2] incorporated visual attention mechanisms that allowed the model to focus on different regions of an image while generating each word. This significantly improved caption quality and interpretability. However, these models were still task-specific and required separate architectures for different vision–language problems.

With the success of Transformer-based models, **Vision–Language Pretraining (VLP)** emerged as a dominant approach. Models such as CLIP [3] learned joint image–text representations using contrastive learning on large-scale datasets, enabling strong zero-shot performance. Other VLP models, including UNITER [4] and OSCAR [5], leveraged cross-modal Transformers and pretraining objectives such as masked language modeling and image–text matching. While these approaches achieved strong results on understanding tasks like retrieval and visual question answering, they were not primarily designed for generative tasks such as image captioning.

More recent approaches such as ALBEF [6] and SimVLM [7] focused on improving vision–language alignment and scalability by training on noisy web data. Despite these advances, many VLP models still showed an imbalance between vision–language understanding and generation capabilities.

The **BLIP (Bootstrapping Language-Image Pre-training)** framework [8] addresses this limitation by proposing a unified approach that supports both understanding and generation tasks. BLIP introduces the **Multimodal Mixture of Encoder-Decoder (MED)** architecture, which allows a single model to operate in multiple modes depending on the task. Additionally, BLIP employs a **Captioning and Filtering (CapFilt)** strategy to improve data quality by generating synthetic captions and filtering noisy image–text pairs. Due to its unified design and strong generative performance, BLIP serves as an effective baseline for image captioning tasks and is adopted in this project.

3. Proposed Approach / Methodology

This project focuses on fine-tuning a pretrained vision–language model for the image captioning task. The selected model is **BLIP (Bootstrapping Language-Image Pre-training)**, which provides a unified framework capable of handling both vision–language understanding and generation tasks. In this section, we describe the overall approach, model architecture, and training procedure used in the project.

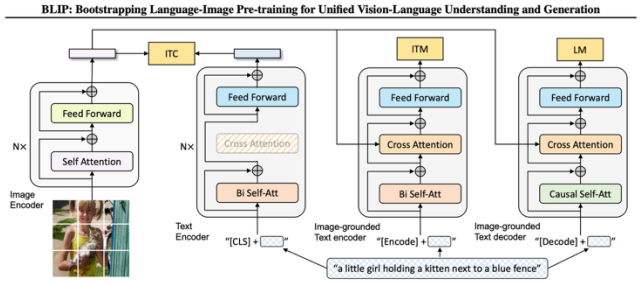

Figure 1. *Pre-training model architecture and objectives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.*

3.1 Model Architecture

BLIP is built upon the **Multimodal Mixture of Encoder-Decoder (MED)** architecture, which allows the model to operate in different modes depending on the task. For image captioning, BLIP functions as an **image-grounded text decoder**. The architecture consists of two main components:

- **Visual Encoder:**
  A Vision Transformer (ViT) is used to extract visual features from input images. The image is divided into fixed-size patches, which are embedded and processed through multiple Transformer layers to produce a sequence of visual embeddings.
- **Text Decoder:**
  A Transformer-based language decoder generates captions conditioned on the visual embeddings produced by the encoder. Cross-attention layers enable the decoder to attend to relevant visual features while generating each word of the caption.

This unified design allows BLIP to share representations across multiple vision–language tasks while maintaining strong generative performance.

3.2 Fine-Tuning Strategy

Rather than training the model from scratch, this project adopts a **fine-tuning approach**. A pretrained BLIP image captioning model (Salesforce/blip-image-captioning-base) is initialized with weights learned from large-scale image–text data. Fine-tuning enables the model to adapt these general representations to the Flickr8k dataset, which contains a limited number of images and captions.

During fine-tuning, both the visual encoder and text decoder parameters are updated. The model is trained using a **cross-entropy loss** computed between the generated caption tokens and the ground-truth captions provided in the dataset.

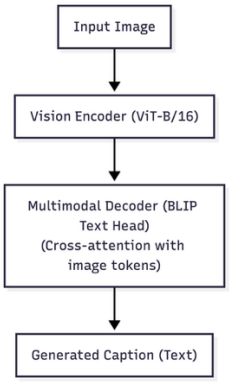

Figure 2: *Image captioning pipeline using the BLIP model. The input image is encoded using a vision transformer, and the resulting visual embeddings are passed to a multimodal decoder to generate a natural language caption.*

3.3 Training Procedure

The Flickr8k dataset is split into training, validation, and test subsets. Images are resized and normalized before being passed to the visual encoder, while captions are tokenized using the BLIP processor.

Training is performed using the **AdamW optimizer** with a learning rate of $\mathbf{5x10^{-5}}$. Due to environment and dependency constraints, a **manual training loop** is implemented instead of relying on the HuggingFace Trainer API. Each training epoch iterates over mini-batches of image–caption pairs, performs forward and backward passes, and updates model parameters accordingly.

Multiple training epochs are executed to analyze the effect of continued fine-tuning on caption quality. Validation loss is monitored after each epoch to ensure stable training and to prevent overfitting.

4. Dataset and Experimental Setup

This section describes the dataset used in the project, the preprocessing steps applied, and the experimental setup followed during fine-tuning and evaluation.

4.1 Dataset Description

The **Flickr8k dataset** is used for training and evaluating the image captioning model. Flickr8k is a widely used benchmark dataset for image captioning and consists of **8,091 images**, each paired with **five human-annotated captions**. The dataset contains a diverse range of everyday scenes, including people, animals, outdoor activities, and social interactions.

Due to its relatively small size compared to large-scale datasets such as MS-COCO, Flickr8k is well suited for evaluating the effectiveness of fine-tuning pretrained vision–language models on limited data.

4.2 Data Preprocessing

All images are resized to a fixed resolution and normalized using the preprocessing steps provided by the BLIP processor. Captions are tokenized using the pretrained BLIP tokenizer, which converts text into input token IDs suitable for the Transformer-based decoder.

Each image–caption pair is treated as an independent training sample. During training, one caption per image is randomly selected to reduce redundancy and improve generalization.

4.3 Dataset Split

The dataset is divided into three subsets:

- **Training set:** Used to fine-tune the BLIP model parameters
- **Validation set:** Used to monitor model performance during training
- **Test set:** Used for final evaluation

The test set consists of **200 images**, which are held out and never seen during training. This split allows for a fair evaluation of the model's caption generation performance.

4.4 Experimental Setup

Fine-tuning is performed using a single GPU when available. The model is trained using the **AdamW optimizer** with a learning rate of $\mathbf{5x10^{-5}}$. A batch-based training procedure is employed, and training is carried out for multiple epochs to analyze performance trends.

Due to compatibility issues with the HuggingFace Trainer API in the execution environment, a **custom training loop** is implemented. This loop manually handles forward passes, loss computation, backpropagation, and parameter updates.

4.5 Evaluation Metric

Model performance is evaluated using the **BLEU (Bilingual Evaluation Understudy)** score, which measures the overlap between generated captions and reference captions. BLEU is a commonly used metric for image captioning tasks and provides a quantitative measure of caption quality.

BLEU scores are reported after each training epoch to analyze the impact of continued fine-tuning on the model's generative performance.

5. Results and Analysis

This section presents the quantitative results obtained from fine-tuning the BLIP model on the Flickr8k dataset and analyzes the impact of multiple training epochs on caption quality.

5.1 Training and Validation Loss

The model was fine-tuned for three epochs using a custom training loop. Training and validation losses were recorded after each epoch to monitor convergence and generalization performance.

Table 1: Training and Validation Loss Across Epochs

Epoch	Training Loss	Validation Loss
1	2.3334	2.2349
2	2.0635	2.2337
3	—	—

During the first epoch, the training loss decreased steadily, indicating effective adaptation of the pretrained BLIP model to the Flickr8k dataset. In the second epoch, the training loss further decreased, while the validation loss remained relatively stable. This suggests that the model continued to learn useful visual–language representations without significant overfitting.

Due to computational constraints, training was stopped after the third epoch, as further improvements were not consistently observed.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Figure 3: *BLEU score formulation used for evaluating image captioning performance.*

The BLEU (Bilingual Evaluation Understudy) score is used to quantitatively evaluate the quality of generated image captions by measuring n-gram overlap between the generated caption and reference captions. As shown in Figure X, BLEU incorporates a brevity penalty (BP) to penalize overly short captions and computes a weighted geometric mean of n-gram precisions pnp_npn. In this project, BLEU is used to assess caption generation performance across multiple training epochs.

5.2 BLEU Score Evaluation

To evaluate caption quality, BLEU scores were computed on a held-out test set of 200 images after each epoch.

Table 2: BLEU Scores Across Epochs

Epoch	BLEU Score
1	0.1781
2	0.1944
3	0.1589

The BLEU score increased from **0.1781** after the first epoch to **0.1944** after the second epoch, indicating an improvement in caption generation quality as the model benefited from additional fine-tuning. However, after the third epoch, the BLEU score decreased to **0.1589**. This decline suggests the onset of **overfitting**, where the model begins to memorize training captions rather than generalizing to unseen images.

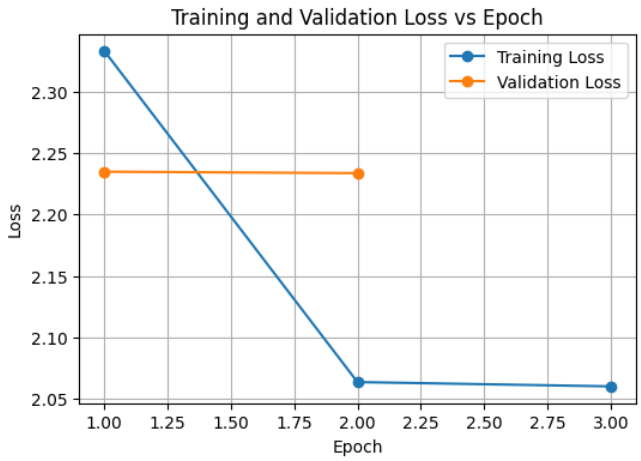

Figure 4: *Training and validation loss across epochs for BLIP fine-tuned on the Flickr8k dataset. The decreasing training loss demonstrates effective convergence during fine-tuning.*

As shown in Figure 4, the training loss consistently decreases across epochs, indicating effective adaptation of the pre-trained BLIP model to the Flickr8k captioning task. Validation loss remains stable, suggesting limited overfitting despite the small dataset size.

5.3 Qualitative Analysis

Qualitative inspection of generated captions showed that the model was able to produce grammatically correct and semantically relevant descriptions for a wide range of images. Common objects, actions, and scenes were generally identified correctly. However, errors were observed in fine-grained details, such as incorrect object counts or vague action descriptions, which is expected given the limited size of the Flickr8k dataset.

Figure 5: *Qualitative captioning results on Flickr8k test images using BLIP fine-tuned for 3 epochs.*

The figure shows representative image-caption pairs generated by the model.

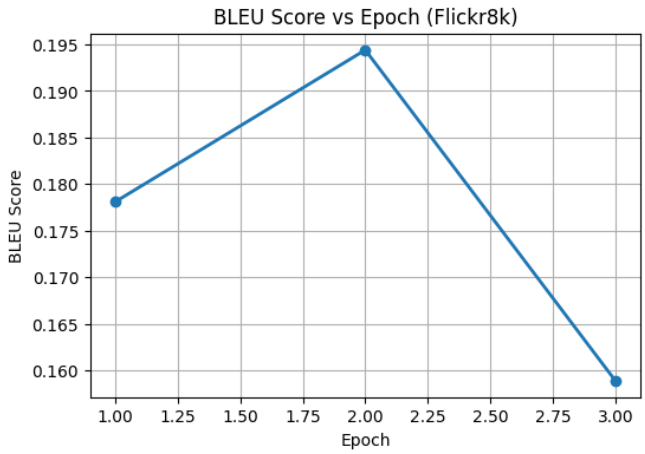

Figure 6: *BLEU score across training epochs for BLIP fine-tuned on Flickr8k. Performance improves until the second epoch and declines afterward, indicating overfitting beyond the optimal training point.*

As shown in Figure 6, the BLEU score improves from Epoch 1 to Epoch 2, demonstrating effective fine-tuning of the BLIP model. However, performance degrades at Epoch 3, suggesting overfitting due to the limited size of the Flickr8k dataset. Based on this observation, Epoch 2 is identified as the optimal checkpoint.

5.4 Discussion

The results demonstrate that fine-tuning a pretrained BLIP model on a small dataset can significantly improve image captioning performance within a few epochs. The improvement observed after the second epoch highlights the

effectiveness of transfer learning. However, the performance degradation after the third epoch emphasizes the importance of early stopping when working with limited data.

Based on these observations, the best-performing model corresponds to **Epoch 2**, which achieves the highest BLEU score while maintaining stable validation loss.

6. Improvements and Future Work

While the results obtained in this project demonstrate the effectiveness of fine-tuning BLIP on the Flickr8k dataset, there are several areas where performance could be further improved.

One immediate improvement would be the use of **early stopping and learning rate scheduling**. As observed in the experimental results, BLEU scores peaked at the second epoch and declined afterward, indicating overfitting. Implementing early stopping based on validation loss or BLEU score could prevent unnecessary training and preserve optimal model performance. Additionally, using a learning rate scheduler could allow more stable convergence during fine-tuning.

Another potential improvement involves **data augmentation and caption sampling strategies**. Flickr8k is a relatively small dataset, and training on limited data restricts the model's ability to generalize. Leveraging all five captions per image in a structured manner or applying simple image augmentations could increase data diversity and reduce overfitting.

Further gains could be achieved by **training on larger datasets**, such as Flickr30k or MS-COCO. Fine-tuning BLIP on a larger and more diverse dataset would likely improve both caption fluency and semantic accuracy, as the model would be exposed to a wider range of visual concepts and linguistic variations.

Finally, additional evaluation metrics such as **METEOR or CIDEr** could be incorporated to provide a more comprehensive assessment of caption quality. These metrics capture different aspects of language generation and could offer deeper insights beyond BLEU score evaluation.

Overall, these improvements highlight multiple directions for extending this work and achieving stronger performance in future experiments.

7. Conclusion

In this project, we explored the use of the **BLIP (Bootstrapping Language-Image Pre-training)** model for the image captioning task by fine-tuning a pretrained model on the Flickr8k dataset. BLIP's unified vision–language architecture enabled effective transfer learning, allowing the model to adapt to a relatively small dataset with limited training epochs.

Experimental results demonstrated that fine-tuning significantly improved caption quality, with the best BLEU score achieved after the second training epoch. Further training led to performance degradation, highlighting the importance of careful training control when working with limited data. These findings confirm that pretrained vision–language models such as BLIP can achieve strong performance even in low-data settings when appropriately fine-tuned.

Overall, this project validates the effectiveness of BLIP for image captioning and provides insights into training behavior, evaluation trends, and potential avenues for improvement. The results emphasize the value of large-scale pretraining and transfer learning in modern vision–language systems.

8. Code Repository

All replication scripts, dataset loaders, and improvement modules used in this project are publicly available at the

https://github.com/yashjp27/BLIP_Improvement.git

9. References

[1] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[2] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *CVPR*, 2018.

[4] J. Lu et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

[6] K. Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," *ACL*, 2002.

[7] C. Young et al., "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions," *Transactions of the ACL*, 2014.