Course Project: CSE 597

Archisman Ghosh
CSE Department
Penn State University
State College, PA, USA
apg6127@psu.edu

***Abstract*—This study focuses on implementing and fine-tuning BLIP (Bootstrapped Language-Image Pretraining), a state-of-the-art model for vision-language tasks. BLIP leverages a Multimodal Mixture of Encoder-Decoder (MED) architecture to effectively integrate image and text representations, excelling in tasks like image-text retrieval and captioning. We reproduce the original results on the Flickr30K dataset and further fine-tune the model to improve performance. Our fine-tuned version outperforms the pre-trained model in several key metrics, including R@1, R@5, and R@10, demonstrating better retrieval accuracy. The findings highlight the effectiveness of fine-tuning BLIP for task-specific improvements and underline its potential for broader vision-language applications**

***Index Terms*—Image-Text Retrieval, Computer Vision, BLIP**

I. Introduction

Image-text retrieval is a key task in computer vision and natural language processing that focuses on aligning visual and textual information to enable seamless cross-modal search. It encompasses two sub-tasks: text-to-image retrieval, where the goal is to find images matching a textual query, and image-to-text retrieval, which identifies textual descriptions corresponding to a given image. This task is vital for applications like visual search, multimedia retrieval, and automated image captioning. The primary challenge lies in bridging the semantic gap between images and text, as they are fundamentally different data modalities. Recent advancements, particularly in deep learning and multimodal pretraining techniques, have led to significant progress. By training on large-scale datasets, modern models, such as those based on transformers, can map both modalities into a shared feature space, achieving state-of-the-art results on benchmarks like Flickr30k and MSCOCO.

A. Challenges in Image-text retrieval

Image-text retrieval faces several challenges due to the fundamental differences between visual and textual data. A primary issue is the **semantic gap**: images and text are inherently dissimilar, with images being high-dimensional arrays of pixel values and text represented as sequential word embeddings. Bridging this gap requires models to capture both low-level features (e.g., objects, colors) and high-level semantics (e.g., relationships, context) effectively. Another significant challenge is **data variability and ambiguity**. Images can contain multiple objects and complex scenes, making it difficult to identify which elements correspond to the text. Similarly, textual descriptions can be ambiguous, abstract, or subjective, with variations in how different annotators describe the same image. This issue becomes more pronounced when dealing with diverse datasets. The **lack of labeled data** for multimodal tasks also poses a bottleneck. While large-scale datasets like Flickr30k and MSCOCO provide valuable benchmarks, they are limited in size and domain diversity compared to what is needed for robust generalization. Furthermore, effective **multimodal alignment** demands sophisticated architectures that can jointly process visual and textual data. Ensuring these models remain computationally efficient while capturing meaningful relationships adds another layer of complexity. Despite recent advances, challenges like bias in data, domain adaptability, and interpretability continue to hinder widespread adoption in real-world scenarios.

Code for this study can be found here.

B. Motivation

Despite advancements in image-text retrieval, existing methods often rely on large-scale annotated datasets, limiting their generalizability to diverse real-world scenarios. Moreover, many models struggle with effectively aligning complex visual and textual semantics, particularly when data is noisy or sparse. To address these limitations, Bootstrapped Language-Image Pretraining (BLIP) [1] introduces an innovative approach that leverages both supervised and unsupervised learning to bridge the semantic gap between images and text. BLIP utilizes a unified vision-language architecture with a novel bootstrapped learning strategy, allowing it to learn from both labeled and unlabeled data. This significantly enhances its ability to generalize across diverse datasets. By incorporating multimodal pretraining and leveraging transformers for cross-modal alignment, BLIP achieves state-of-the-art performance in retrieval tasks. Its robust design and pretraining strategies make BLIP a compelling solution for overcoming the challenges of semantic alignment, data efficiency, and scalability in image-text retrieval tasks.

C. Contributions

The major contributions of this study are as follows:

1) We reimplement the code for BLIP and reproduce the results with a comparison with the state-of-the-art methods on the Flickr30k dataset.

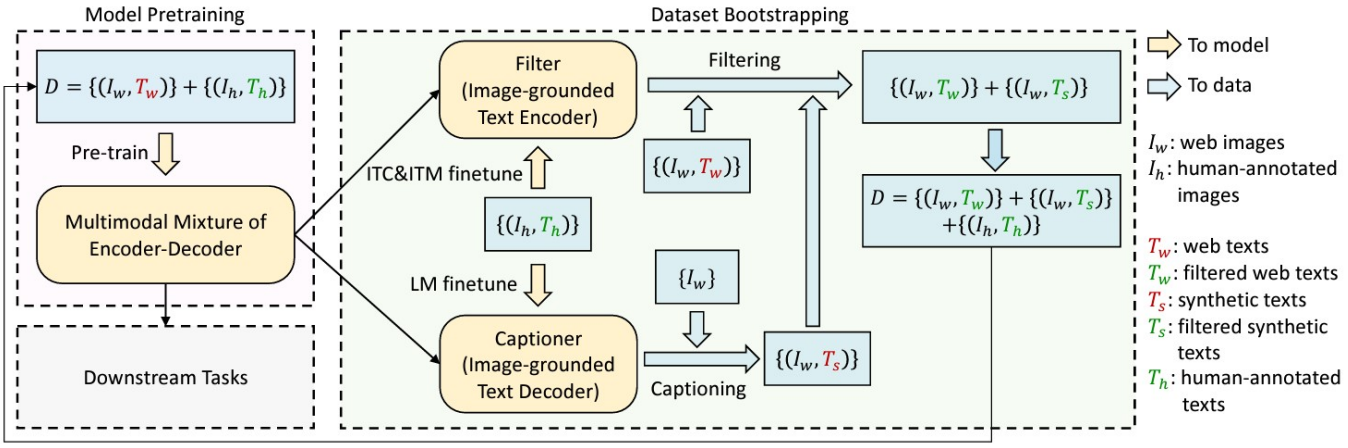

Fig. 1. Learning framework of BLIP [1]. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The bootstrapped dataset is used to pre-train a new model.

2) We further propose an improvement over the original idea.

D. Report Structure

Section II provides a background and related works. Section III covers the dataset. Section IV describes the procedure and architecture of the neural network classifier. Section V presents the results. We discuss the results in Section VI and conclude in Section VII.

II. Background

A. BLIP

Bootstrapped Language-Image Pretraining (BLIP) is a unified framework for multimodal tasks such as image-text retrieval, utilizing a transformer-based architecture to align visual and textual representations effectively. It combines a visual encoder, a text encoder, and a multimodal fusion module to map images and text into a shared embedding space. BLIP employs a bootstrapped learning strategy, integrating supervised learning from annotated data and self-supervised learning from unannotated data, enhancing its scalability and generalization capabilities. To optimize its performance, BLIP uses multiple loss functions, including contrastive loss for aligning image-text pairs in a shared space, image-text matching loss to determine pair relevance, and language modeling loss for generating textual descriptions. These components allow BLIP to capture complex cross-modal relationships and achieve state-of-the-art results in image-text retrieval, particularly in diverse and challenging datasets.

B. Competing methods and ideas

The performance of BLIP is compared with the following competing methods.

1) UNITER: It focuses on learning joint representations using an encoder-based model to align visual and textual modalities effectively [2]. It achieves impressive recall rates of 65.7% (R@1), 88.6% (R@5), and 93.8% (R@10) for text retrieval, demonstrating strong performance in matching image-text pairs and excelling in understanding multimodal relationships.

2) VILLA: It employs adversarial training to improve the robustness of vision-language models [3]. It leverages perturbations in multimodal embeddings during pretraining to make models less sensitive to input noise. This approach enhances fine-grained alignment and generalization across diverse datasets, making VILLA a reliable framework for multimodal tasks.

3) OSCAR: It enhances alignment between visual and textual features by using object tags as anchor points during pretraining [4]. It achieves notable recall rates of 70.0% (R@1), 91.1% (R@5), and 95.5% (R@10) for text retrieval, effectively integrating object-level semantics into multimodal representations for improved retrieval performance.

4) UNIMO: UNIMO integrates cross-modal contrastive learning to address challenges in unified-modal understanding and generation [5]. It emphasizes end-to-end pretraining on both visual and textual modalities to improve semantic alignment and adaptability across various vision-language processing tasks, showing versatility in handling diverse datasets.

5) ALIGN: It scales up vision-language representation learning using noisy text supervision from large web datasets [6]. Its design achieves remarkable results, including a recall rate of 77.0% (R@1) in text retrieval, highlighting its capability to handle massive, uncurated data and learn robust multimodal embeddings.

6) ALBEF: It emphasizes aligning vision and language representations before fusion to enhance retrieval performance [7]. It achieves recall rates of 77.6% (R@1), 94.3% (R@5), and 97.2% (R@10) for text retrieval, using a combination of

pretraining objectives like contrastive and image-text matching loss for effective multimodal learning.

III. Architecture of BLIP

A. Multimodal Mixture of Encoder-Decoder

BLIP introduces MED (Multimodal Mixture of Encoder-Decoder) architecture designed for flexible and efficient multimodal learning. MED comprises:

1) Visual Encoder: A Vision Transformer (ViT) extracts patch-level image features, encoding each image as high-dimensional embeddings. Let an image I be processed into patch embeddings $\mathbf{v} \in \mathbb{R}^{N \times d}$, where N is the number of patches and d is the embedding size.

2) Text Encoder: A transformer-based language model encodes input text into embeddings $\mathbf{t} \in \mathbb{R}^{L \times d}$, where L is the sequence length.

3) Fusion model: A transformer fuses $\mathbf{v}$ and $\mathbf{t}$ into a unified embedding space $\mathbf{z}$, facilitating both understanding tasks like retrieval and generation tasks like captioning.

MED supports Encoder-only, Decoder-only, and Encoder-Decoder modes for finer-grained multimodal tasks.

B. Pre-training Objectives

1) Image-Text Contrastive Learning: The contrastive loss aligns image and text embeddings in a shared feature space. For an image-text pair $(\mathbf{v}_i, \mathbf{t}_i)$, the loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(\mathbf{v}_i,\mathbf{t}_i))}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{v}_i,\mathbf{t}_j))} + \log\frac{\exp(\text{sim}(\mathbf{t}_i,\mathbf{v}_i))}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{t}_i,\mathbf{v}_j))}\right]$$

where $\text{sim}(\cdot)$ is a similarity function, such as cosine similarity.

2) Image-Text Matching: This binary classification task predicts whether an image-text pair is semantically related. The loss is computed as:

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

where $y_i \in \{0,1\}$ is the ground truth label, and p_i is the predicted probability.

3) Image-Conditioned Language Modeling: This generative objective models the likelihood of a text sequence $\mathbf{t}$ given an image $\mathbf{v}$. The loss is defined as:

$$\mathcal{L}_{\text{LM}} = -\sum_{j=1}^{L}\log P(t_j|\mathbf{v}, t_{<j})$$

where t_j is the j-th token in the text sequence.

C. Caption Filtering

Caption Filtering (CapFilt) is a key pretraining strategy in BLIP designed to handle the challenges of noisy web-crawled data by generating and refining high-quality pseudo-captions. Using BLIP's multimodal encoder-decoder architecture, pseudo-captions are generated to provide relevant textual descriptions of images. These captions are then evaluated through a filtering mechanism based on factors like semantic alignment between image and text embeddings, linguistic fluency, and contextual relevance. Low-quality captions are removed, ensuring that only accurate and meaningful image-text pairs are retained. This iterative process progressively refines the dataset, enabling the model to learn robust representations from large-scale, uncurated data without being hampered by noise. By reducing dependence on manual annotations and improving data quality, CapFilt empowers BLIP to generalize effectively across vision-language tasks, achieving superior performance in image-text retrieval, captioning, and other multimodal applications.

IV. Results

A. Dataset

In the context of this report, the Flickr30K dataset is chosen for the finetuning of the BLIP model for Image-Text Retrieval. Adhering to the paper's methodology, the Karpathy split is adopted, distributing the Flickr30K dataset into 29,000 training images, 1,000 images for validation, and 1,000 images for the testing phase. This extensive collection serves as a pivotal training ground for enhancing the model's proficiency in correlating visual content with relevant textual descriptions.

TABLE I
BLIP vs SOTA

Method	# Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	87.3	98.0	99.2	75.6	94.1	96.8
VILLA	4M	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR	4M	-	-	-	-	-	-
UNIMO	5.7M	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN	1.8B	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF	14M	95.9	99.8	100.0	85.6	97.5	98.9
BLIP	14M	96.6	99.8	100.0	87.2	97.5	98.8

B. Comparison with SOTA

The performance improvements brought by BLIP are evident in Table I, highlighting its significant advancements over previous methods. With the same set of 14 million pre-training images, BLIP outperforms the former leading model, ALBEF, by +2% in average recall@1 on the Flickr30K dataset. Furthermore, a zero-shot retrieval evaluation was performed by applying the model fine-tuned on COCO directly to the Flickr30K dataset. Results, as detailed in Table II, reveal a substantial margin of superiority over prior approaches. The evaluation focuses on standard information retrieval metrics: recall@1, recall@5, and recall@10. Recall@1 assesses how

often the most relevant result appears at the top of the ranking, while R@5 and R@10 measure the effectiveness of retrieving relevant results within the top 5 and top 10 positions, respectively. These metrics are essential for evaluating a model's ability to handle larger retrieval tasks and accurately identify captions or descriptions corresponding to specific images.

TABLE II
BLIP VS SOTA FOR ZERO-SHOT IMAGE-TEXT RETRIEVAL

Method	# Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.5	100.0	84.9	96.7	98.3

C. Reproducing results

To provide an apple-to-apple comparison, we implement the BLIP as is on the Flickr30K dataset, and compare it with the baseline, which is the results reported in Table I. We observe in Table III that we obtain comparative results on reproducing the results with the original weights in the pre-trained model.

TABLE III
REPRODUCED RESULTS FOR BLIP ON FLICKR30K DATASET

Method	Flickr30K (1K test set)					
	TR			IR		
	R@1	R@5	R@10	R@1	R@5	R@10
BLIP (reproduced)	97.0	99.8	100.0	87.1	97.6	98.76
BLIP (baseline)	96.6	99.8	100.0	87.2	97.5	98.8
BLIP (fine-tuned)	96.2	99.8	100.0	84.9	98.6	98.3

D. Fine-tuning

Following up on the official documentation of the project, we change the hyperparameters in the `retrieval_flickr.yml` file to optimize the model to use Image-Text Contrastive (ITC), and Image-Text Matching (ITM) loss, we set the value of k to 128 and reduce the train and test batch sizes to 8 and 32 respectively. The fine-tuning has been done on a single NVIDIA 3090 GPU. We further observe the results in Table III.

V. DISCUSSION

In this section, we discuss the results observed due to the fine-tuning of the BLIP model.

A. Text Retrieval

In the Text Retrieval (TR) scenario, the fine-tuned model we developed, performs slightly worse in R@1 compared to the model with the original weights and the one reported in the original paper. However, it still demonstrates strong accuracy, achieving perfect scores in both R@5 and R@10. This indicates that nearly all relevant texts are retrieved within the top 10 results, showing that the model is highly effective in broader retrieval tasks, even though it lags slightly in ranking the most relevant text at the top.

B. Image Retrieval

In the Image Retrieval (IR) scenario, there is a noticeable difference in the R@1 score. The fine-tuned model has a lower ability to retrieve the most relevant image on the first attempt compared to both the replicated results and the model from the original paper. Nevertheless, it performs well in R@5 and R@10, suggesting that it is still proficient at retrieving relevant images within the top 10 results, albeit with some room for improvement in pinpointing the most relevant image as the top result. The replicated model with original weights outperforms the fine-tuned model in both TR and IR tasks across all metrics, closely followed by the results reported in the original paper. This highlights that while the fine-tuning process improved the model, the original weights remain slightly superior, particularly in ranking the most relevant result.

C. Effect of Batch Size

Owing to memory and time constraints, we had to reduce the batch size to fine-tune the model. This likely contributed to a performance dip.

1) Critique: Smaller batch sizes can lead to less accurate gradient estimates, potentially causing suboptimal convergence. While they may improve generalization by introducing noise and preventing overfitting on smaller datasets, this noise can hinder learning on larger, more complex tasks. Larger batches, by contrast, smooth the learning process and act as a form of regularization. However, smaller batches can introduce noise that helps escape local minima, though excessive noise may prevent convergence, especially on complex tasks with a rugged loss landscape.

2) Suggested hyperparameter changes: Reducing the learning rate (possibly to half) and increasing the number of epochs (by 2 or 3) can help mitigate the negative effects of smaller batch sizes by allowing the model to make more gradual updates and refine its weights over time. This approach improves convergence, enhances stability, and helps the model better navigate the loss landscape. These could not be implemented and tested owing to time and resource constraints but may aid in the improvement of the overall performance of the model.

VI. CONCLUSION

This study reproduces the results performed by BLIP (Bootstrapped Language- Image Pretraining) on the Flickr30K dataset and further fine-tunes the model for better performance. The reproduced results are discussed and potential hyperparameter changes to workaround the diminishing performance of the fine-tuned model are also suggested.

REFERENCES

[1] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.

[3] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning, 2020.
[4] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020.
[5] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning, 2022.
[6] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.
[7] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021.