CAP 6412 Assignment 1: BLIP-2

Zhen Hao Sia
zhenhao.sia@ucf.edu

Abstract

All inference scripts for BLIP-2 [2] are written from scratch with reference to code from the original LAVIS repository; none of the original inference scripts in the original repository are used directly to replicate results for Task A.

Newton is not used when running the inference scripts; all scripts are ran on a personal machine with Ubuntu 22.04, 32GB RAM and one RTX 4090 with 24GB vRAM.

Instructions to setup the inference scripts and datasets are outlined in the README.md file provided with the source code. All computed results are stored in the source code to avoid recomputation every time inference is ran.

Source code:
https://github.com/siatheindochinese/cap6412_assign1

1. Task A

This section covers the replication of results outlined in the original BLIP-2 paper.

1.1. BLIP-2 Stage 1 model fine-tuned on coco

Replication of results is done on COCO and Flickr30K using BLIP-2's QFormer finetuned on the COCO dataset.

Following section 4.4 of BLIP-2 paper, only image-text pairs in the top k image-text contrastive (ITC) scores for image-to-text and text-to-image are used before computing image-text matching (ITM). For both image retrieval and text retrieval, reranking is done on the ITM scores respectively to determine Recall@1, Recall@5 and Recall@10.

Addendum: k is set to 64 instead of 128 due to time and computational constraints.
Addendum 2: Only 1000 images out of 5000 images in the COCO test set is used due to time and computational constraints.

Metrics (%)	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
Flickr30K	97.7	99.9	99.9	89.7	98.1	98.9
COCO	93.6	99.6	99.8	83.4	96.8	99.8

1.2. Captioning and VQA with OPT6.7B model

For the captioning task, all 5000 images in the COCO test set are used.

For the visual question answering (VQA) task, only 50,000 out of 200,000 questions are evaluated due to time and computational constraints.

$pycocoltools$, $pycocoevalcap$, vqa and vqa_eval included in the LAVIS library/repository are used to extract performance metrics.

Captioning Results:

Metrics	CIDEr	BLEU@4
Captioning	144.963	43.330

VQA Results:

Metrics	Accuracy (%)
VQA	31.03

2. Task B

This section seeks to evaluate the VQA performance of BLIP-2 when extended to videos instead of images, in particular for the videos in the MSVD-QA dataset. Only the validation dataset of MSVD-QA (6000+ questions) is used due to time and computational constraints.

Evaluation of generated answers is treated as a multimodal classification task; each postprocessed result (after processing punctuations and digit articles) is compared with the exact ground truth answer, following the evaluation scheme in the MSVD-QA paper, "Video Question Answering via Gradually Refined Attention over Appearance and Motion" [1] and code from $lavis/tasks/multimodal\_classification.py$ and $lavis/tasks/vqa.py$.

2.1. Randomly Sample 1 frame from each video

1 frame is randomly sampled from a video and VQA is done using said frame as an image input.

MSVD-QA Results:

Accuracy (%)	2.14

2.2. Uniformly Sample 10 frames from each video

10 frames are uniformly sampled from a video. Each frame is processed by BLIP-2's image encoder, and the resulting embeddings for each frame are temporally average-pooled before being passed into QFormer along with the text input.

Accuracy (%)	0.00

3. References

[1] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y. (2017, October). Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1645-1653).

[2] Li, J., Li, D., Savarese, S., Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.