NTU AI 2024 HW1 Report

B10902037 Yu Xiang Luo

Task 1

1. Briefly describe how you implement the two models:

   In this assignment, I utilize the huggingface packages to conduct the experiments. Below is the main package I use:

   - `transformer.AutoModel...` → Load models.
   - `datasets.load_dataset` → Load datasets.
   - `torch.utils.data.DataLoader` → Process data in batch to inference faster.
   - `evaluate` → Corpus BLEU, ROUGE, and METEOR function.

   Also, I convert the captions to lowercase and remove punctuations for better evaluations.

2. Metrics Result:

	MSCOCO-Test				flickr30k			
	BLEU	ROUGE-1	ROUGE-2	METEOR	BLEU	ROUGE-1	ROUGE-2	METEOR
BLIP	0.3237	0.4023	0.1594	0.2776	0.2398	0.3206	0.1091	0.2033
Phi-4	0.2560	0.3971	0.1517	0.3325	0.2658	0.3636	0.1400	0.3043

3. Analysis:

   - In the MSCOCO experiments, BLIP outperforms Phi-4 on BLEU metrics but is outperformed by Phi-4 on METEOR. This suggests that BLIP's output aligns more closely with the specific wording used by MSCOCO annotators, leading to higher BLEU scores. In contrast, Phi-4, as a larger model, likely generates captions with a more diverse vocabulary, improving its METEOR score by capturing broader semantic similarities.
   - In the Flickr30k experiments, Phi-4 outperforms BLIP across all metrics, indicating that Phi-4 is more effective for this dataset.
   - Comparing the two datasets, we observe that the overall metric scores are higher for MSCOCO, suggesting differences in dataset complexity, with MSCOCO potentially being easier for models to perform well on.