Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis

Priyanshu Sharma (862395994), Sahil Chowkekar (862393156)

March 12, 2023

1 General Descriptions

- The complete code is present in this **github/priyanshu-sharma/VLM**. Presently, it is in a development state, so we keep it in a private repo. But anyone can request access to this repo.
- We use the BART-based model for pre-training our model. In the original paper, they pre-trained the "facebook/bart-base" model using Masked Language Modeling (MLM) and Textual Aspect-Opinion Extraction (AOE) as a Textual Pre-training Task, Masked Region Modeling (MRM) and Visual Aspect-Opinion Generation (AOG) as a Visual Pre-training Task, and Multimodal Sentiment Prediction (MSP) as a Multimodal Pre-training Task.
- Additionally, We have also compared the pre-trained Bart model with three different pooling techniques, which are:-
  - **FIRST** - The first token of the multimodal input sequence is always a weighted sum of the 36 regional image features. Its final hidden state is considered as the aggregate multimodal sequence representation with visual representations as queries.
  - **CLS** - Similarly, the final hidden state for the special token (i.e., [CLS] token in the sentence input) is the aggregate representation with textual representations as queries.
  - **BOTH** - Concatenate above two hidden state at once.
- Overall, we have presented the pre-training of the BART model's results using two sections:-
  - **Evaluation Metrics** - We have compared the five different evaluation metrics for all three pooling techniques, which are: - Dev Evaluation Accuracy, Dev F1 Score, Dev Loss, Dev Evaluation Loss and Dev Global Step.

  - **System Metrics -** These are more related to GPU Performance, such as GPU Memory Allocated, GPU Utilization, etc.

- Contact the authors of this report for more information.

2 Evaluation Metrics

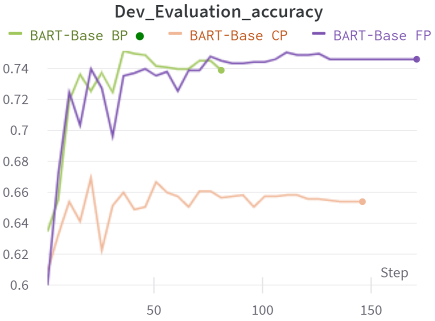

Figure 1: Dev Evaluation Accuracy

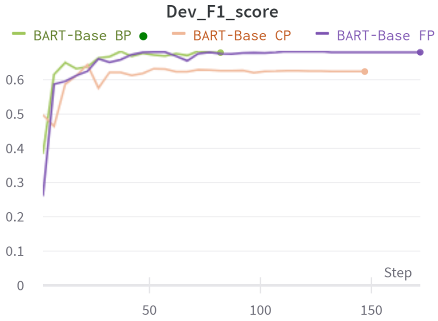

Figure 2: Dev F1 Score

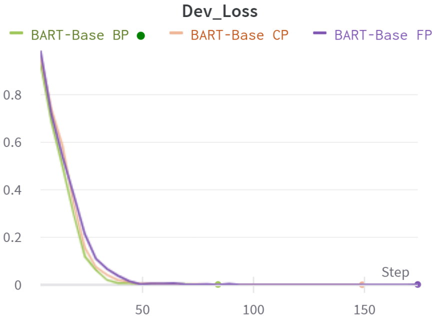

Figure 3: Dev Training Loss

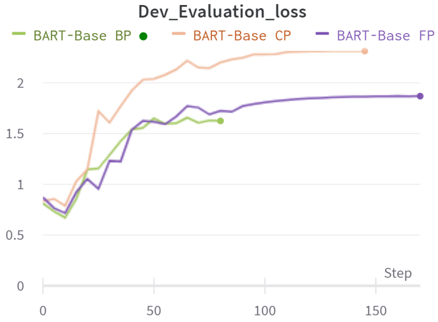

Figure 4: Dev Evaluation Loss

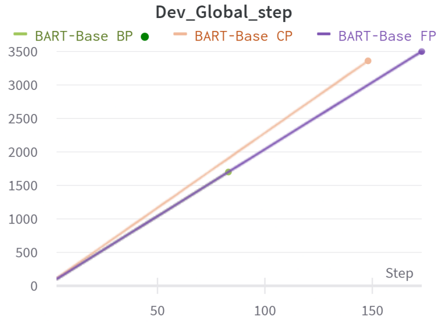

Figure 5: Dev Global Steps

2.1 Results

- For Dev Evaluation Accuracy, we get the best score by BP technique, which is about 75.13%, with FP technique, we get best score around 75.04%, but performance with CP technique suffers, which is around 66.92%.
- BP technique outperforms other pooling technique, with best F1 score of 68.42%, 67.65% with FP technique and 64.42% is best F1 score with CP technique.
- Training Loss is almost comparable for all three different pooling techniques and remains constant after inital 50 steps.
- Overall Bart model with CP technique has higher Evaluation Loss as compared to other techniques as a result it hurts its overall evaluation accuracy.

3 System Metrics

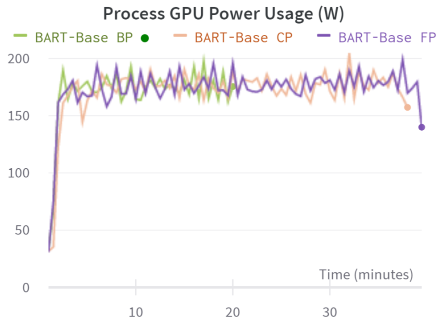

Figure 6: GPU Power Usage in Watts

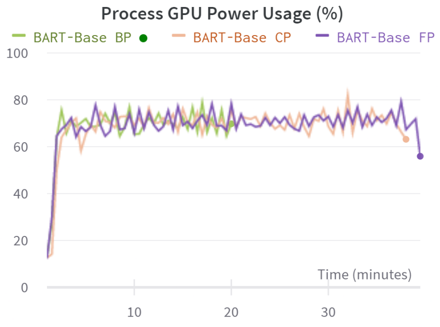

Figure 7: GPU Power Usage in %

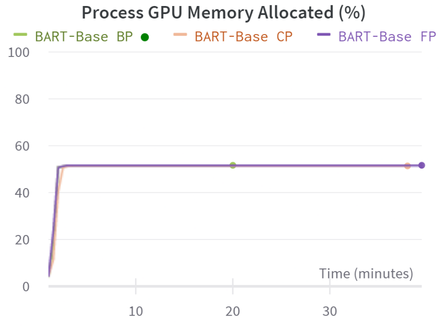

Figure 8: GPU Memory Allocated %

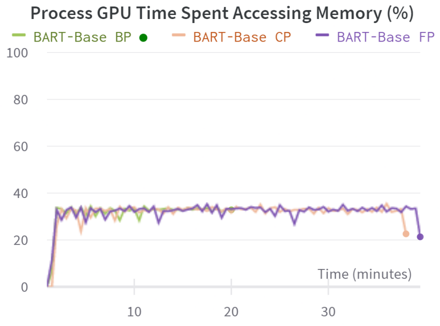

Figure 9: GPU Time Spent Accessing Memory %

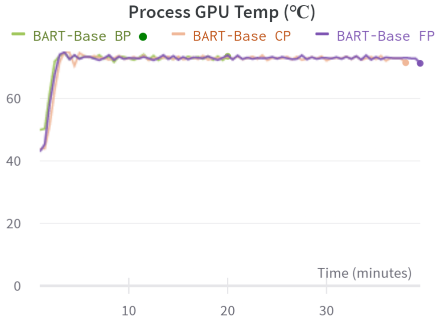

Figure 10: GPU Temperature in C

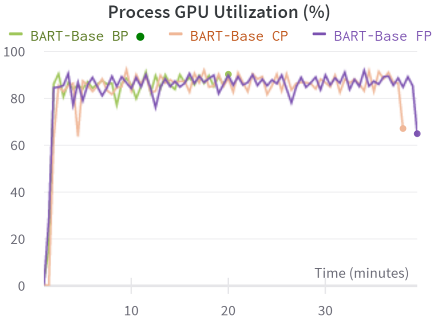

Figure 11: GPU Utilization %

3.1 Results

- Almost all the system-related metrics, such as GPU memory allocated, GPU Temperature, etc, remain similar for all three different techniques.