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 two different datasets, as stated in the original paper, for downstream task. These datasets are downloadable from this **link** and collectively known as **MVSA-Multi** and comprises of "Twitter - 2015" and "Twitter - 2017" datasets.
- We consider all the three subtasks in MABSA as our downstream tasks, including Joint Multimodal Aspect-Sentiment Analysis (JMASA), Multimodal Aspect Term Extraction (MATE), and Multimodal Aspect-oriented Sentiment Classification (MASC).
- We evaluate our model over three subtasks of MABSA and adopt MicroF1 score (F1), Precision (P) and Recall (R) as the evaluation metrics to measure the performance. For MASC, we also compare the Accuracy with other approaches. Overall, for downstream tasks, we presented the results using two sections:-
  - **Evaluation Metrics** - We have compared the five different evaluation metrics for all three pooling techniques, which are: - Dev Recall, Dev MicroF1 Score, Loss, Dev Precision and MASC Accuracy.
  - **System Metrics** - These are more related to GPU Performance, such as GPU Memory Allocated, GPU Utilization, etc.
- Other than evaluating the performance with Facebook's Base BART Model, we also evaluate all the three downstream subtasks on both the dataset with BART Large Model.
- Contact the authors of this report for more information.

2 Evaluation Metrics

2.1 Twitter - 2015 Dataset

2.1.1 MASC

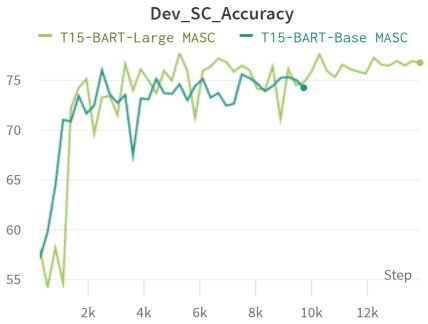

Figure 1: MASC Dev Accuracy

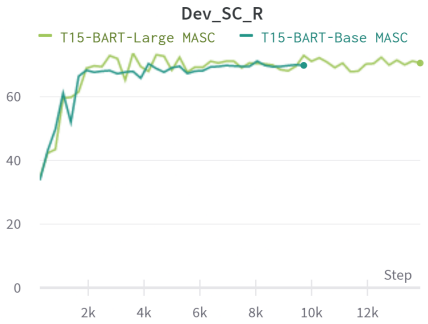

Figure 2: MASC Dev Recall

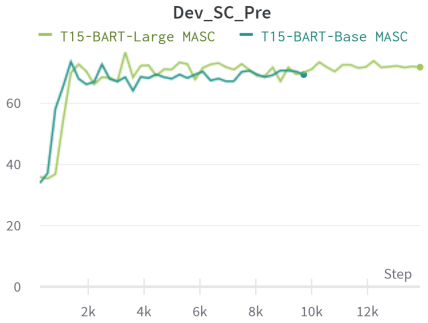

Figure 3: MASC Dev Precision

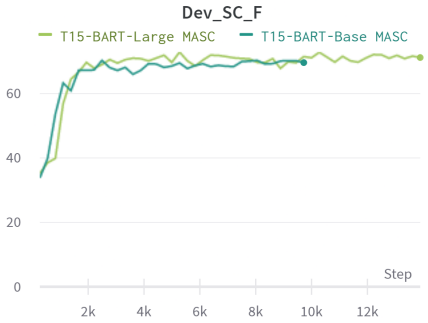

Figure 4: MASC Dev MicroF1

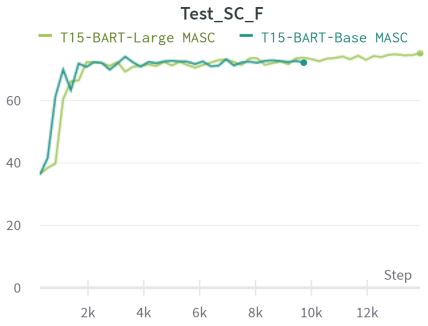

Figure 5: MASC Test MicroF1

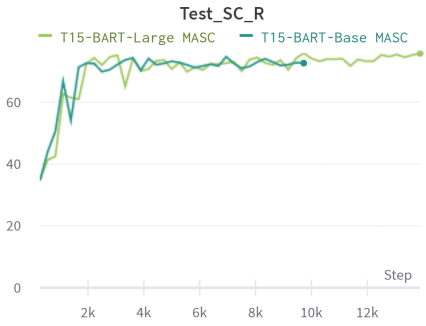

Figure 6: MASC Test Recall

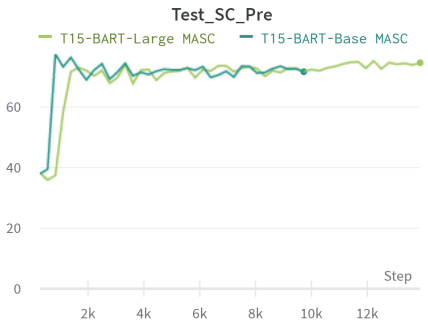

Figure 7: MASC Test Precision

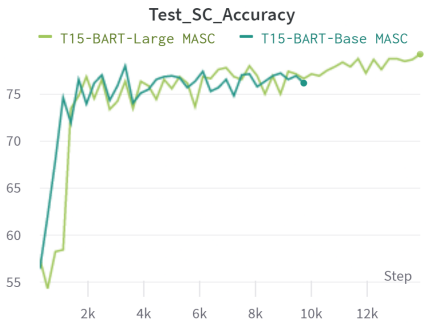

Figure 8: MASC Test Accuracy

2.1.2 MASC + MATE

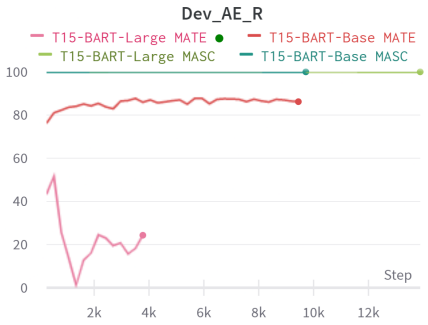

Figure 9: Dev Recall

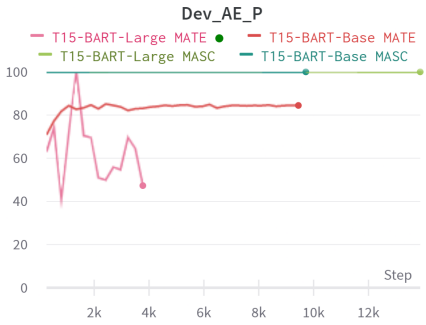

Figure 10: Dev Precision

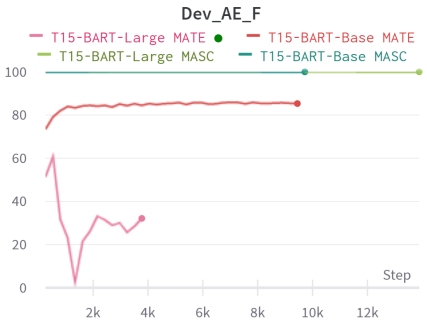

Figure 11: Dev MicroF1

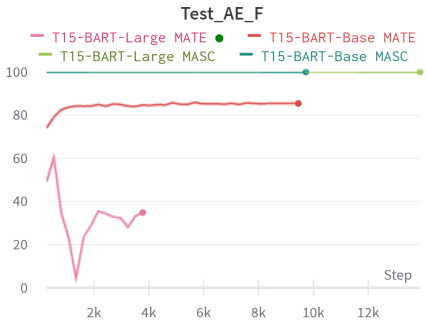

Figure 12: Test MicroF1

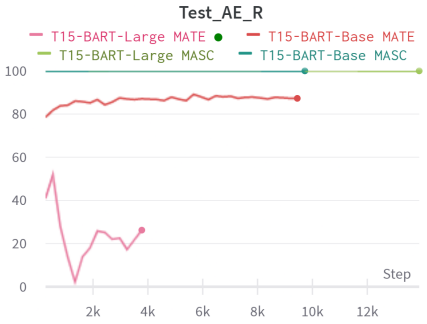

Figure 13: Test Recall

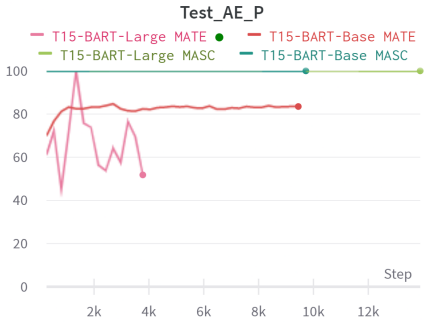

Figure 14: Test Precision

2.1.3 MAESC

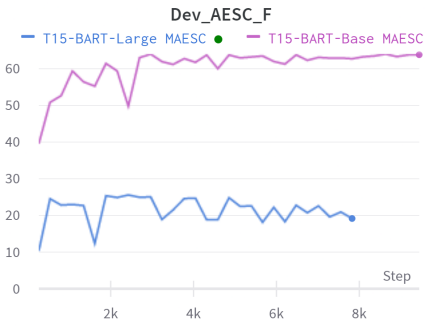

Figure 15: MAESC Dev MicroF1

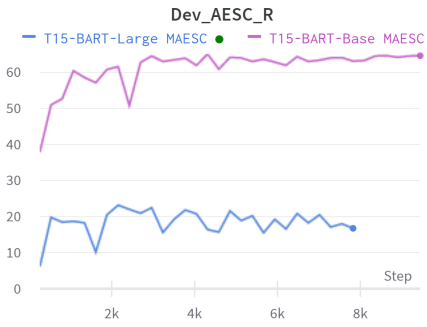

Figure 16: MAESC Dev Recall

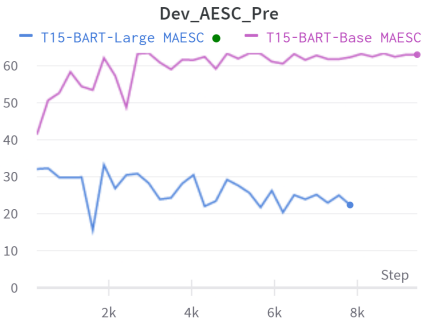

Figure 17: MAESC Dev Precision

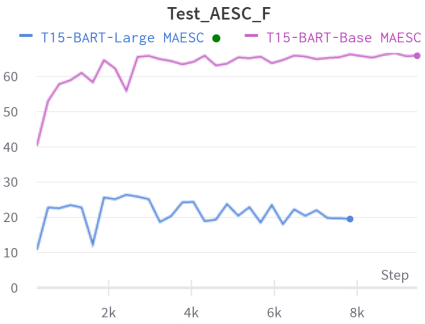

Figure 18: MAESC Test MicroF1

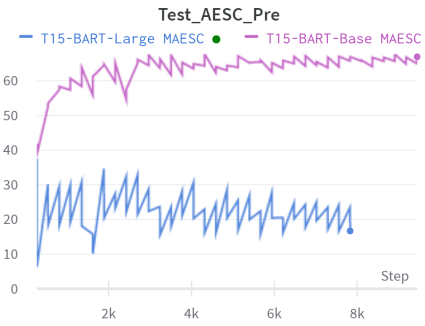

Figure 19: MAESC Test Precision

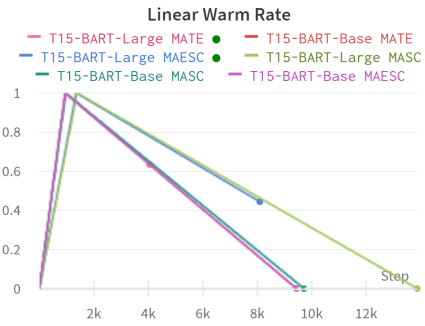

Figure 20: Linear Warm Rate

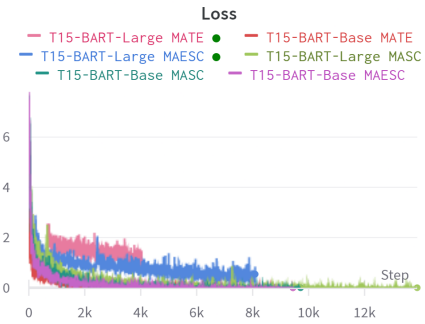

Figure 21: Loss

2.2 Twitter - 2017 Dataset

2.2.1 MASC

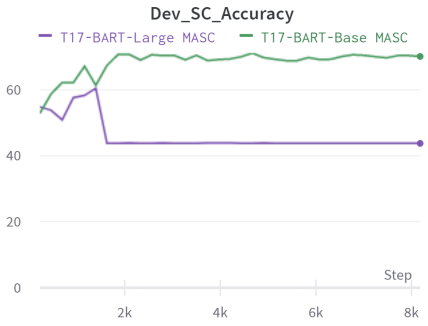

Figure 22: MASC Dev Accuracy

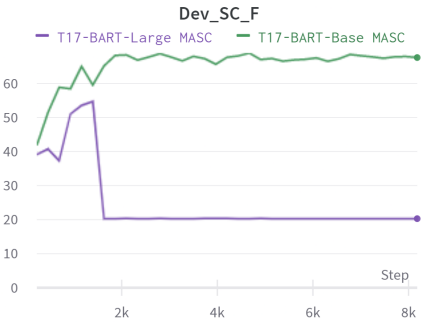

Figure 23: MASC Dev MicroF1

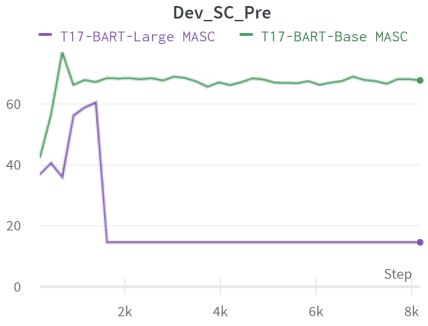

Figure 24: MASC Dev Precision

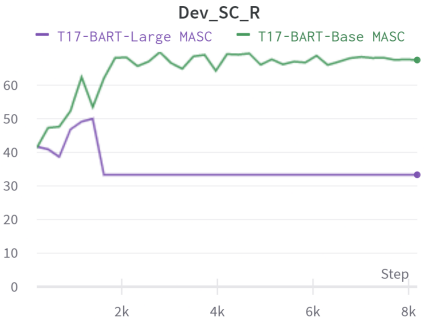

Figure 25: MASC Dev Recall

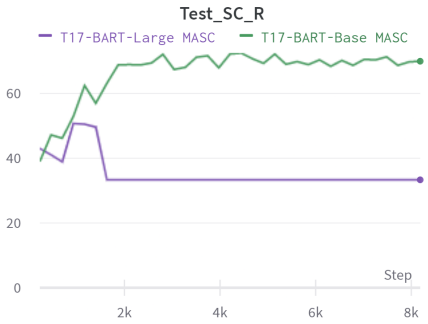

Figure 26: MASC Test Recall

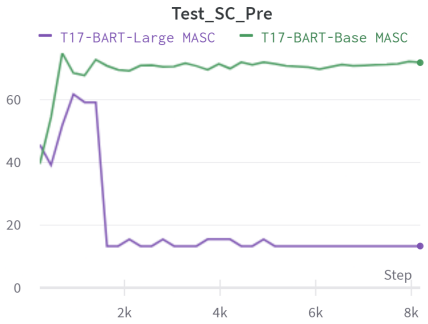

Figure 27: MASC Test Precision

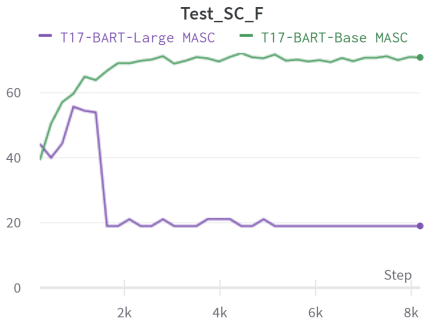

Figure 28: MASC Test MicroF1

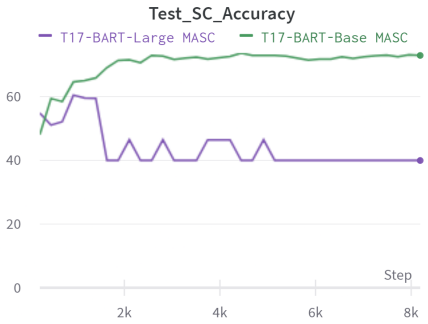

Figure 29: MASC Test Accuracy

2.2.2 MASC + MATE

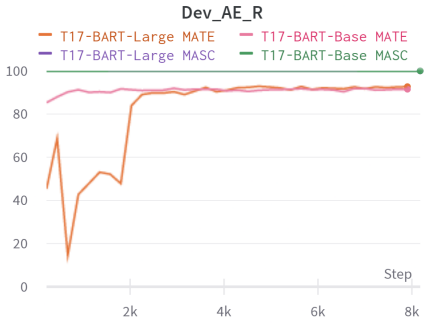

Figure 30: Dev Recall

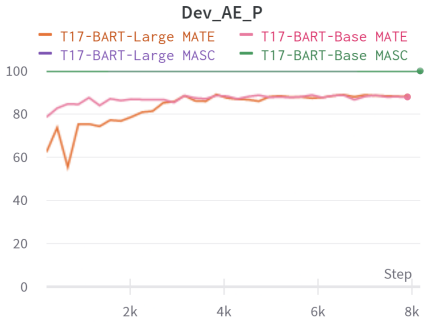

Figure 31: Dev Precision

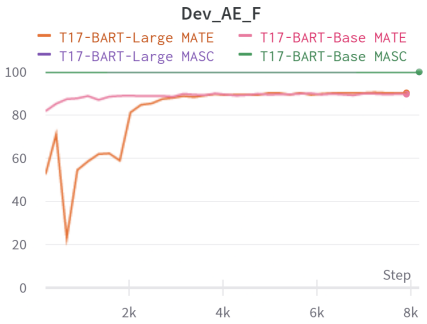

Figure 32: Dev MicroF1

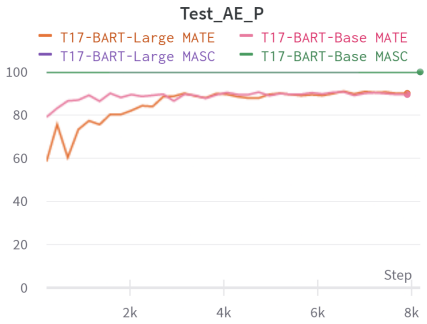

Figure 33: Test Precision

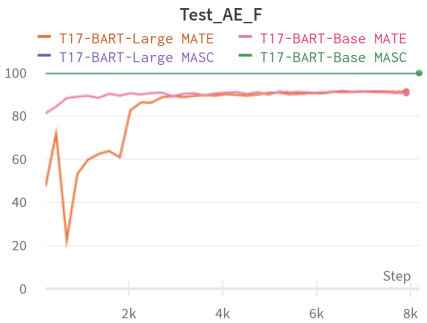

Figure 34: Test MicroF1

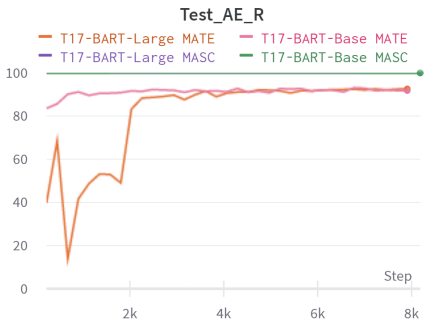

Figure 35: Test Recall

2.2.3 MAESC

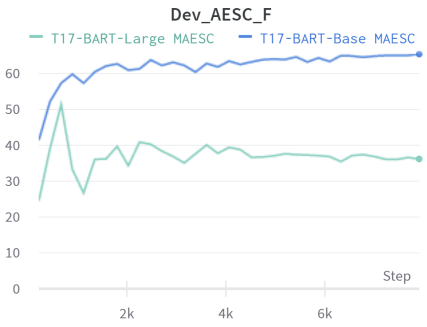

Figure 36: MAESC Dev MicroF1

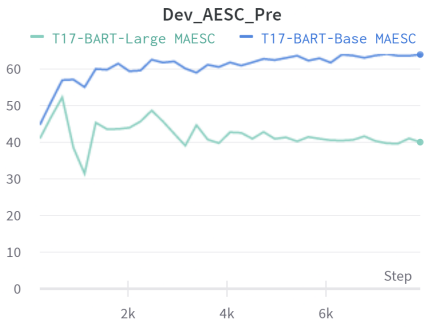

Figure 37: MAESC Dev Precision

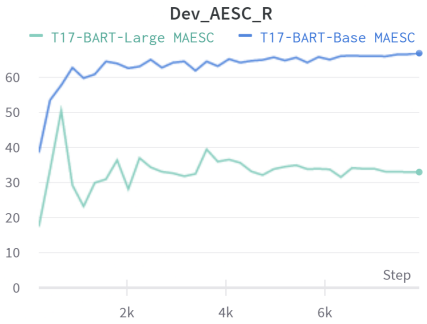

Figure 38: MAESC Dev Recall

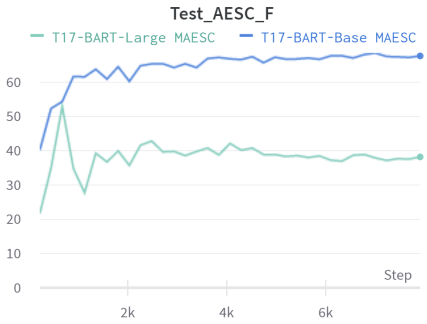

Figure 39: MAESC Test MicroF1

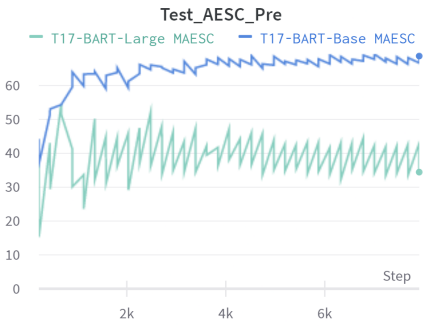

Figure 40: MAESC Test Precision

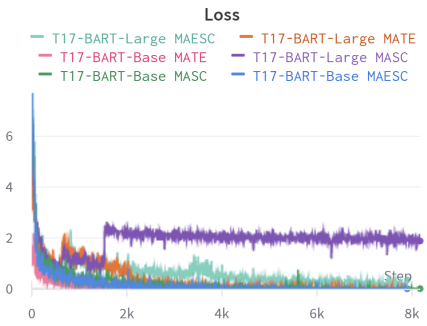

Figure 41: Loss

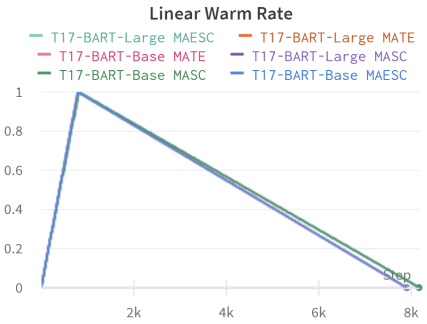

Figure 42: Linear Warm Rate

3 System Metrics

3.1 Twitter - 2015 Dataset

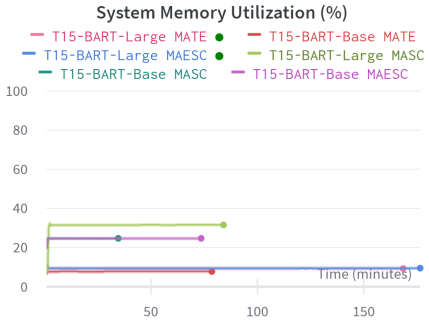

Figure 43: System Memory Utilization %

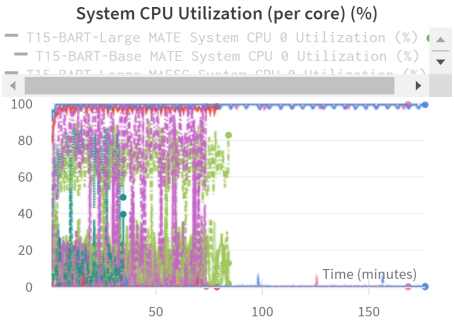

Figure 44: System CPU Utilization %

3.2 Twitter - 2017 Dataset

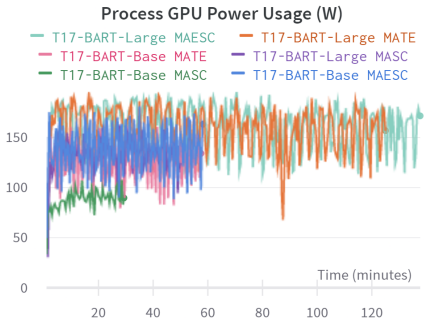

Figure 45: GPU Power Usage in Watts

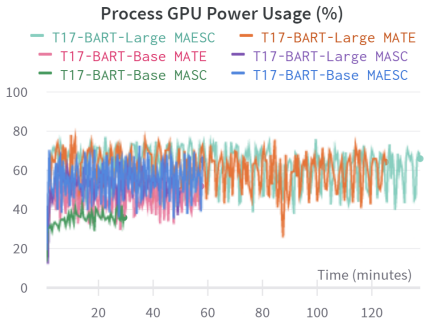

Figure 46: GPU Power Usage in %

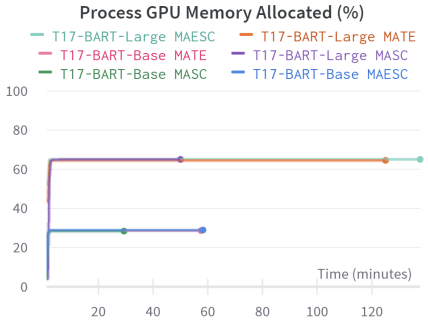

Figure 47: GPU Memory Allocated %

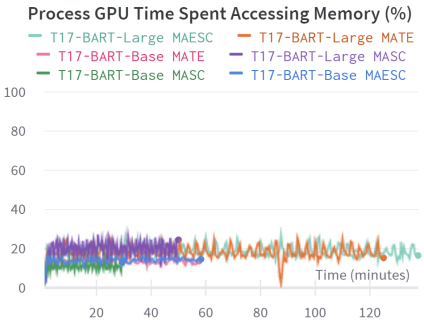

Figure 48: GPU Time Spent Accessing Memory %

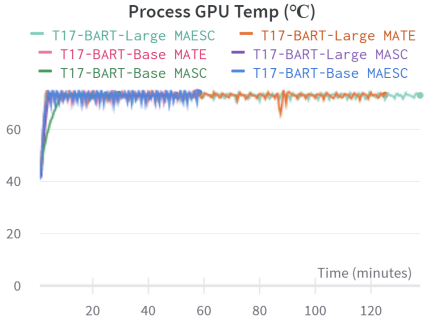

Figure 49: GPU Temperature in C

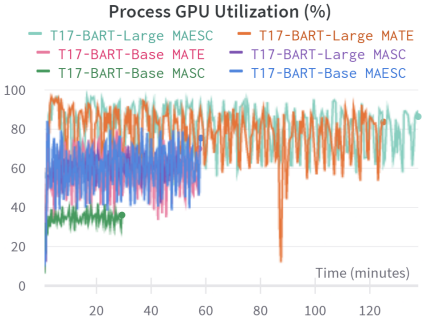

Figure 50: GPU Utilization %

3.3 Results

- For MASC Task, we found that our performance degrade on Twitter - 2017 dataset as we increase the size of model (BART-Large), but boost the performance upto 79.27 % for Test Accuracy Metrics on Twitter - 2015 dataset.
- For MATE Task, Base and Large Model's performance coverages to same value on "Twitter - 2017" dataset, but degrade or fluctuate rapidly with large model on other dataset.
- For MAESC Task, we observed the performance degradation with increase in model size and remains constant (60 % - 65 %) in all metrics with smaller model on both the dataset.
- Overall, we observed the higher memory and GPU utilization for larger Bart model across all the three downstream tasks.