보고서

- 팀원 : 최예원, 전서현
- 성능 (결과 캡쳐) - 1000장으로 test

```
apss017@elogin0:~/apss017_project/why$ ./run.sh -n 1000
srun: job 812445 queued and waiting for resources
srun: job 812445 has been allocated resources

==========================================
 Model: Variational AutoEncoder (VAE)
------------------------------------------
 Validation: OFF
 Number of images: 1000
 Input binary path: ./data/input_fp32.bin
 Model parameter path: /opt/apss24/project/param_fp32.bin
 Answer binary path: /opt/apss24/project/answer_fp32.bin
 Output binary path: ./data/output.bin
==========================================

Initializing input and parameters...Done!
Generating images...Done!
Elapsed time: 41.039582 (sec)
Throughput: 24.366720 (images/sec)
Finalizing...Done!
Saving output to ./data/output.bin...Done!
```

- 최적화 내용 간단 설명

1. lmodel.cu와 Tensor.cu 파일 작업을 통해 GPU에 전체적인 변수를 올리고 연속적인 layer 연산을 GPU에서 진행
2. Linear layer) weight를 전치하는 과정과 input 과 weight의 연산을 타일링으로 구축
3. reshape layer) CPU 성능보다 GPU 성능이 더 낮아 해당 작업은 CPU에서 진행
4. convTransposed2d, conv2d, BatchNorm layer) 다량의 for문을 gpu에서 작업하여 각각의 thread를 분배하여 연산 진행