

Configuration Manual

MSc Research Project MSc in Data Analytics

Sayok Kumar Bose Student ID: X20187688

School of Computing National College of Ireland

Supervisor: Mr. Rejwanul Haque

National College of Ireland





National
College of
Ireland

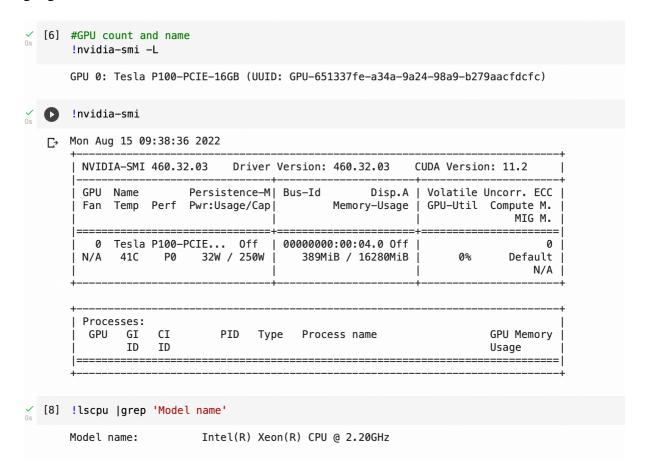
Student Name:	Sayok Kumar Bose		
Student ID:	X20187688		
	MSc in Data Analytics		
Programme:		22	
Module:	MSc Research Project		
Supervisor:	Mr. Rejwanul Haque		
Submission Due Date:	15/08/2022		
Project Title:	Generating Python Code from Docstrings using OpenNMT		
Word Count:	6232 Page Count18		
	ndard specified in the report template. To use other autigal (plagiarism) and may result in disciplinary action. Sayok Kumar Bose	hor's written or	
Signature.			
Date:	15/08/2022		
	FOLLOWING INSTRUCTIONS AND CHECKLIST		
	opy of this sheet to each project (including multiple copies)		
Attach a Moodle su each project (includir	bmission receipt of the online project submission, to na multiple copies).		
You must ensure tl	nat you retain a HARD COPY of the project, both for and in case a project is lost or mislaid. It is not sufficient to		
Assignments that are	e submitted to the Programme Coordinator Office must be plac eed outside the office.	ed into the	
Office Use Only			
Signature:			
Date: Danalty Applied (if an	valicable).		
renaity Applied (If ap	pplicable):		

Configuration Manual

Sayok Kumar Bose Student ID: X20187688

1 Hardware Configuration

The research was conducted over Google Collab Pro sessions with high ram as the selected runtime environment. Here are the following hardware details of the underlying machine in google collab.



2 Software Configuration

For the research OpenNMT toolkit was used installation steps can be easily found here ¹.



¹ https://opennmt.net/OpenNMT-py/main.html#installation

3 OpenNMT

Connecting to google drive:

```
[ ] from google.colab import drive
    import os

os.environ['C /content/drive (cmd + click) = "0"
    drive.mount('/content/drive', force_remount=True)
    os.chdir("/content/drive/MyDrive/research-project/nmt")
```

Configuration for NMT:

```
27 save_model: model/model.code

28
29 # Stop training if it does not imporve after n validations
30 early_stopping: 3
31
32 # Default: 5000 - Save a model checkpoint for each n
33 save_checkpoint_steps: 1000

45 # To save space, limit checkpoints to last n
36 keep_checkpoint: 3
37
  37
38 seed: 3435
39
40 # Default: 100000 - Train the model to max n steps
41 train_steps: 100000
42
43 # Default: 10000 - Run validation after n steps
44 valid_steps: 1000
 44 valid_steps: 1000
45
46 # Default: 4000 - for large datasets, try up to 8000
47 warmup_steps: 4000
48 report_every: 100
49
50 # Activate TensorBoard
51 tensorboard: true
52 tensorboard_log_dir: model/tensorboard
53
  54

55 decoder_type: transformer

56 encoder_type: transformer

57 word_vec_size: 512

58 rnn_size: 512

59 layers: 6

60 transformer_ff: 2048

61 heads: 8
  62
63 accum_count: 1
64 optim: adam
65 adam_beta1: 0.9
66 adam_beta2: 0.998
67 decay_method: noam
68 learning_rate: 2.0
69 max_grad_norm: 0.0
70
   71 # Tokens per batch, change if out of GPU memory
 72 batch_size: 2048
73 valid_batch_size: 2048
74 batch_type: tokens
```

OpenNMT training and translation:

```
**Convert to a few words
# Convert to a few words
# test: head -n 50 data/DataForLMG.test.subword | cut -d" " -f-15 > data/DataForLMG.test.subword.lm

# Generation
ommt_translate -model model/model-lm_step_20000.pt -src data/DataForLMG.test.subword.lm -output data/output.subword -n_best 1 -random_sampling_topp 0.9 -beam_size 10 -gpu
# Desubwording
python3 ~/scripts/desubword.py vocab/source.model data/output.subword
```

Installing SacreBLEU for Evaluation: