National College of Ireland

Configuration Manual

MSc Research Project
Artificial Intelligence

Taylou Maniganze
Student ID: 22162071

School of Computing
National College of Ireland

Supervisor: Prof. Paul Stynes

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Taylou Maniganze
Student ID:	22162071
Programme:	MSCAI1 Jan23 Artificial Intelligence **Year:** 2023
Module:	MSc Research in computing
Lecturer:	Prof. Paul Stynes
Submission Due Date:	27/05/2024
Project Title:	A Deep Learning Framework to Identify Interconnectivity of Citation Networks.
Word Count:	**787 Page Count: 3**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Taylou Maniganze

Date: 27/05/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Taylou Maniganze
Student ID: 22162071

1 Overview

The Deep Learning Framework to identify interconnectivity of citation networks combines transformer models and Interconnectivity. The transformer models are trained on three different datasets C2D-I (Roman et al., 2022), and SciCite. The experiments of this study were performed using pretrained transformer models BERT_base, BERT_large and DistilBERT to identify the model with the highest accuracy to be used for interconnectivity.

The experiments were conducted on two platforms AWS EC2 instance and Kaggle notebook.

2 Kaggle setup

To run the deep learning model on Kaggle you need to have an account with Kaggle and sign in, after signing in go to settings and verify your account as you cannot access the accelerators without verifying your account. After, load the datasets from the upload option then choose the accelerator as TPU VM v3-8 and make sure the internet option is on to allow the download of the libraries.

After the setup go to the code and change the file path (file_path = '/kaggle/input/ric-experiments/C2D-I.csv') to your Kaggle path. Before running the script run the following:

- **pip install nltk:** to be able to use the natural language toolkit.
- **pip install transformers:** to be able to use the hugging face transformer models (huggingface.co, 2024).

The above libraries should allow the code to run smoothly if not then install the rest of the libraries manually **pip install scikit-learn** and **pip install seaborn**.

The TPU is free, it provides 20 hours per week to use it but because of its popularity sometimes there is a queue, and you just have to wait for your turn. The waiting time depends on your number in the queue.

On Kaggle the script will take around 30 minutes with sciCite dataset and 1 hour and 30 minutes for the C2D-I dataset.

3 EC2 instance setup

To run the deep learning model on the AWS EC2 instance I used the account provided by the college and the following were the settings:

- This study referred to these steps Set up to use Amazon EC2 - Amazon Elastic Compute Cloud to set up the instances. I started working with G4dn.4xlarge but the model needed more memory I switched to G4dn.12xlarge with 8 GPUs.
- After the instance has been created, it needs to be configured to serve the study's purpose, the study followed these steps Configure your Amazon Linux instance - Amazon Elastic Compute Cloud
- Because I was using a Linux instance to access it, I used SSH and depending on the instance needs for the model Tensorflow needs NVIDIA libraries to speed up the process the libraries needed are Compute Unified Device Architecture (CUDA) and Cuda Deep Neural Network (cuDNN). The libraries have different versions depending on one's SSH kernel's version and to download the libraries you need an account with NVIDIA. The libraries can be found here CUDA Installation Guide for Linux (nvidia.com).
- Transferring datasets and other files to the instance can be done through any FTP application, in these experiments FileZilla was used.
- To run the script Jupyter Notebook was used in the SSH run **sudo apt install python3** then Jupyter Notebook it shall open with the files in it.

On the instance the models will take 15-30 minutes. Since all the models used the same code only a few lines were modified to avoid creating many scripts with the same code. To change the model uncomment the specific line and comment the line above **# tokenizer = AutoTokenizer.from_pretrained('bert-base-uncased')**
 # tokenizer = AutoTokenizer.from_pretrained('distilbert-uncased')

To change the dataset since SciCite has all columns named this needs to be commented **df.columns = ['SrNo', 'TextWhereRefMention', 'Category'],** then Uncomment these lines as you comment the lines above them **# df['CategoryEncoded'] = encoder.fit_transform(df['target']), # df['text'] = df['text'].apply(preprocess_text), # filtered_indices = df['text'].apply(lambda x: len(tokenizer.tokenize(x))) <= max_len, # encoded_texts = tokenizer(filtered_df['text'].tolist(), padding=True, truncation=True, max_length=max_len, return_tensors='tf')**.

To test the model's capabilities to predict if a text from a paper containing citations or reference belongs to any of the three categories background, method or results the code has where to input text and it predicts which category it belongs. When the prompt comes copy from a database that is being used and press enter it should show 0, 1 or 2 as output.

References

Configure your Amazon Linux instance - Amazon Elastic Compute Cloud (no date). Available at: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Configure_Instance.html (Accessed: 25 April 2024).

CUDA Installation Guide for Linux (no date). Available at: https://docs.nvidia.com/cuda/cuda-installation-guide-linux/index.html (Accessed: 25 April 2024).

google-bert (BERT community) (2024). Available at: https://huggingface.co/google-bert (Accessed: 25 April 2024).

Set up to use Amazon EC2 - Amazon Elastic Compute Cloud (no date). Available at: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/get-set-up-for-amazon-ec2.html (Accessed: 24 April 2024).

Roman, M., Shahid, A., Khan, S., Koubaa, A. and Yu, L. (2021). Citation Intent Classification Using Word Embedding. IEEE Access, 9, pp.9982–9995. doi:https://doi.org/10.1109/access.2021.3050547.