Coursework 3 – Deep Learning

Deprivation, or the "the damaging lack of material benefits considered to be basic necessities in a society", is a major concern in modern societies. While many times confused with poverty, in reality, it is a complex, multidimensional, social phenomenon that can severely impact the quality of life of those who are deprived. For banking, Financial Inclusion, or the area of finance focused on giving access to financial services to those who, mostly due to being deprived in different areas, is a key focus of modern regulatory efforts. In this coursework, we will use our knowledge of deep learning to better understand how living conditions in an area, as shown by LIDAR images, can impact deprivation indexes. The seven deprivation indexes that are commonly measured for a geographical area are:

- Income: Measures the proportion of the population experiencing deprivation relating to low income.
- Employment: Measures the proportion of the working-age population in an area involuntarily excluded from the labour market.
- Education: Measures the lack of attainment and skills in the local population.
- Health: Measures the risk of premature death and the impairment of quality of life through poor physical or mental health.
- Crime: Measures the risk of personal and material victimization at the local level.
- Barriers to Housing: Measures the physical and financial accessibility of housing and local services.
- Living Environment: Measures the quality of both the indoor and outdoor local environment.

On the other hand, LiDAR[1] is a technology that uses laser pulses to create 3D and elevations renderings of objects and terrains, similar to how radar technology uses sound. Due to its low cost and efficiency in representing elevations, several governments around the world have collected this data and made it freely available for their citizens to use. As an interesting source of unstructured data, in this coursework, we will explore its societal use to predict multidimensional deprivation. In particular, the UK government has made available both borough-level data regarding multidimensional deprivation and accurate elevation data at 1m spatial resolution for all of England.

You are given a dataset composed of the geographic information of a specific area covering parts of London, UK. There are two parts of the data: images of the neighbourhood (unstructured data) and embedding data, which contains the geographic information of each neighbourhood (structured data). For the embedding data, please refer to the data description file. As for the image data, please note that the format of the image name is "'LIDAR'+'ID'+.'png'". Under this format, you can connect the image with specific geographic information contained in the embedding data.

You can download the images you need from this link: https://uwoca-my.sharepoint.com/:u:/g/personal/cbravoro_uwo_ca/Ea8hL1Qqz-1DqXPUkFg3_OkBkT_oOJ5EdvwX1YU_afWF1w?download=1. It is highly recommended you

[1] More information on LiDAR: https://en.wikipedia.org/wiki/LIDAR and the UK Gov's LiDAR open data initiative https://data.gov.uk/dataset/f0db0249-f17b-4036-9e65-309148c97ce4/national-lidar-programme

download this link directly to your Google Drive (via Colab) instead of downloading and uploading to the cloud. Also, keep this in your Google Drive and just mount it, so you don't download it every time you run your code. To download directly use:

!wget "https://uwoca-my.sharepoint.com/:u:/g/personal/cbravoro_uwo_ca/Ea8hL1Qqz-1DqXPUkFg3_OkBkT_oOJ5EdvwX1YU_afWF1w?download=1"

In your Colab session. Then move the file to your Google Drive.

You will focus your analysis on just one of the dimensions of the index. Please register which one you will work on by including your student ID in the following file https://uwoca-my.sharepoint.com/:x:/g/personal/cbravoro_uwo_ca/ES2EleYnh69Dq2kyMVlxOaEB7wOLEmEXbwTqkwf-3Jw4Rw?e=5aRHAS

Using these data and your knowledge from the course, write a report answering the following questions:

1. (10%) Discuss what relations you think you will find (if any) between your chosen index and the LiDAR images. Support your findings with relevant literature.
2. (40%) Using the images as input and your chosen index as output, calibrate a neural network using **two** models from **different families (for example, ResNet and VGG)** from the Keras Application Library. Two models are in different families if they are not from the same overall architecture. For example, VGG16 and VGG19 would be in the same family, while VGG19 and ResNet150 would not. Train your models setting more than just the dense layers as trainable (so you must train some non-dense layers) discussing the reasoning for your choice. Compare the results from the two models using a reasonable error measure. How do the performances compare? Why?
3. (20%) Choose the model with the highest performance according to question 2 and apply GradCAM to it over a 10-picture sample, chosen so that they represent the whole range for the index. What information is the model using to make the prediction? What conclusions can you make? Discuss your outputs and the model's effectiveness.
4. (20%) What can you say about the use of LiDAR data to measure multidimensional deprivation? Do you think it is a useful tool? Do you foresee any privacy or ethical challenges? If so, how would you propose overcoming them? Make an argument, supported by literature if necessary, regarding the appropriateness of LiDAR data for this topic.

The remaining 10% is given by the format and style as discussed in the rubric.

Conditions of the coursework

Software: You must use Python to run the numerical calculations over your portfolio. A copy of your jupyter notebook must be attached to the coursework as an appendix in a readable format, and a link to the notebook must also be included. Instructions on how to export to PDF can be found here: https://stackoverflow.com/questions/52588552/google-co-laboratory-notebook-pdf-download. The notebook text MUST be machine-readable (so no screenshots of the notebook please) **otherwise a 25% discount will apply**.

Word Limit: 2000 words +/-10% either side of the word count is deemed to be acceptable. Any text that exceeds an additional 10% will not attract any marks. The relevant word count *includes* items such as cover page, executive summary, title page, table of contents, tables, figures, in-text citations and section headings if used. The relevant word count *excludes* your list of references and any appendices at the end of your coursework submission (including the code).

You should always include the word count (from your software word processor, not Turnitin), at the end of your coursework submission, before your list of references.

Title/Cover Page: You must include a title/ cover page that includes: your Student ID, Course Code, Assignment Title, Word Count. This assignment will be marked anonymously, please ensure that your name does not appear on any part of your assignment **otherwise a discount will be applied**.

Submission Deadline: December 17th, 23:59. **Due to internal university deadlines this date cannot be postponed.**

Turnitin Submission: The assignment MUST be submitted electronically via OWL. All required papers may be subject to submission for textual similarity review to the commercial plagiarism detection software under license to the University for the detection of plagiarism. All papers submitted for such checking will be included as source documents in the reference database for the purpose of detecting plagiarism of papers subsequently submitted to the system. Use of the service is subject to the licensing agreement, currently between The University of Western Ontario and Turnitin.com (http://www.turnitin.com).

Late Submission: Late submissions are possible up to seven days after the deadline. There is a linear 10% penalty per day of late submission (Final mark = Original mark – 10% * day) subtracted directly from the final mark. Submissions after the seven days are not accepted and will be considered a non-submission.