Assignment 3: Tutorial | GRAD-E1394

1. **Tentative tutorial title:**

   Topic Modelling with Deep Learning: Policy Applications

2. **Names of all group members. Also enter these on Moodle.**

   Carmen Garro (226594)

   Maria Fernanda Ortega Valencia (229757)

   Gresa Smolica (224725)

3. **Description of preferred topic: Up to 300 words detailing what your tutorial will focus on, including resources you will use to develop it.**

   Aim of the tutorial:

   The aim of the tutorial is to present the application of deep learning techniques for advanced topic modelling in policy analysis. Policy analysis and deep learning have become increasingly interconnected, with the latter opening up new options for extracting subtle insights from complicated textual data.

   Background and motivation:

   Traditional topic modelling techniques, such as Latent Dirichlet Allocation (LDA), have been critical in finding salient topics in policy debates. However, their shortcomings in dealing with contextual nuances, and latent structures in policy-related text, have cleared the way for the incorporation of deep learning approaches. When mixed together with topic modelling, deep learning approaches, such as Neural Topic approaches, show great performance in collecting detailed policy-related themes (Zhao. et al, 2021) since they use neural network topologies to discover hidden patterns and connections within the data (ibid).

   Main contents of the tutorial:

   The tutorial will focus on Transformer-based NLP topic modelling, and will provide a walk-through of neural network architectures suitable for topic modelling. Among the key components building the tutorial will be: discussion on data-collection, preprocessing adapting to policy-contexts and jargons, fine-tuning pre-trained models based on policy-oriented datasets, and show-case real world examples of deep learning applications in public policy.

   Supporting resources:

   - Literature: Some (but not only) of the guiding research papers will be:
     - "Pattern Recognition and Machine Learning", by Christopher M. Bishop
     - "Speech and Language Processing", by Daniel Jurafsky and James H. Martin

    - "Topic Modelling Meets Deep Neural Networks: A Survey", by He Zhao, et al.
    - "Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media", by Patrick Jansson and Shuhua Liu
- Libraries: Open-source libraries such as TensorFlow and PyTorch will be used to supplement the practical parts, so as to provide hands-on experience in developing and fine-tuning neural topic models.
- Datasets: Publicly available datasets from different policy domains will be analysed during this period, to ensure the tutorial is applicable across a wide range of policy situations.

4. **Description of alternative topic: Up to 200 words detailing what your tutorial will focus on in case you are not able to work on your preferred topic, including resources you will use to develop it.**

Aim of the tutorial:

Achieving super performing deep neural networks comes at the cost of training the model for thousands of hours on special units such as GPU-s. Considering that the development of deep learning will not stop, the tutorial seeks to provide some "green" ways of computation that can achieve high performance while not harming the environment.

Background and motivation:

Despite deep learning's exponential advancement and application in almost all spheres of life, there is a growing concern about its associated environmental costs, due to the high computational energy they require. Research shows that the model training is not the only part that requires caution and computation efficiency. Once the model is trained, it will be implemented and used. This power hungry-process called inference takes place continuously and can therefore exceed the energy consumption of the training after a certain number of inference events (Mehlin, et al, 2023). Introducing efficient computation techniques through the tutorial brings us closer to a measurable impact toward the mitigation of climate change.

Main contents of the tutorial:

The tutorial will present some energy-efficient techniques when working with image data and CNN-s. Among the key components building the tutorial will be: discussion on data-collection, comparisons with conventional not-efficient techniques and hands-on exercises where the differences between techniques can be spotted.

Supporting resources:

- Literature: Some (but not only) of the guiding research papers will be:
    - "Towards energy-efficient Deep Learning: An overview of energy- efficient approaches along the Deep Learning Lifecycle", by Vanessa Mehlin, Sigurd Schacht, Carsten Lanquillon

    - “Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models”, by Lasse F. Wolff Anthony, Benjamin Kanding, Raghavendra Selvan
- Libraries: Open-source libraries such as TensorFlow and PyTorch will be used to supplement the practical parts, so as to provide hands-on experience in developing and fine-tuning the CNN-s.
- Datasets: Publicly available image datasets will be analysed during this period and the dataset most suitable for these tasks will be selected for application.

5. Optional comments, such as why your group wants to focus on that topic.

As elaborated above.