DS425 – 2023-2024 projects

Leveraging Large Language Models (LLMs) to accelerate analytics workflows.

The projects being undertaken this year are closely related to operational reality, and the work you will be performing here has the potential to significantly influence and assist (BEL) operations in theatres all over the world. These projects are supported by multiple partners.

We'll provide you a crash course on the fundamentals and the latest advances of natural language processing (NLP), and we'll also talk about how it may be extended to multimodal LLMs to handle audio, images, or video in many contexts and languages.

Together with the other subjects covered in the course, you should use these strategies to offer a proof of concept for accelerating information processing times, extracting key information components, classifying data, and other tasks.

These projects are going to really help you in understanding what goes on inside LLMs; you will learn how to apply various methods, familiarize yourself with various machine learning frameworks, and gain an understanding of how your thoughts at the RMA can be extremely helpful for operational requirements.

You are expected to work in groups of 3 to 4. A report outlining your findings is required of you, and it must be handed in the week preceding the exam. This report counts as one of the tests. The report should at least cover your problem definition, the methods, data, results and some reflections on performance and scalability. There will also be (a) question(s) on the exam related to the project.

Context

From an intelligence standpoint, open data sources have shown to be extremely helpful in recent conflicts (such as those between Israel and Palestine and Ukraine and Russia). Data acquired from these sources can be used to evaluate troop movements, measure damage, make precision strikes, and get insights into the prevalent narratives that exist among various demographic groups. Nevertheless, due to personnel and time restrictions, it is not possible to cover all the material or extract all the information because of the sheer volume of data, which is frequently multimodal (text, images, and videos) and spans multiple languages. This is the point at which you enter the scene. It is possible to mitigate or address some of these issues by using various AI approaches.

Data collection as such is not a problem for your projects because high quality data, of which a significant portion has gone through some stage of the intelligence cycle, may be supplied through a web endpoint (specifics ULT).

Applications of interest

Below you find a (non-exhaustive) list of possible applications. Any other ideas are welcome but should be validated by the course staff first.

- Extracting the content of videos in a low-resource language (e.g. Wolof) and subsequently extracting the topics or identifying the presence of specific narratives.
- Annotations of images that can be used for downstream tasks (beyond named entity recognition), such as clustering, recommendations ((approximate) nearest neighbor search), classification.
- Using image embeddings directly for downstream tasks (cf. supra).
- Text extraction from images for downstream tasks (cf. supra).
- Object detection and classification pipelines for specific tasks (e.g. destroyed equipment, counting equipment instances etc.)
- Few-shot or zero-shot classification of text.

Problem specific resources:

Below you will find a non-exhaustive list of models that are suited for the applications in the different project.

- Video and or audio to text:
  - Whisper (OpenAI) – includes seamless translation.
  - Seamless Communication (Meta) – the medium version supports Wolof, quality of the results TBD.
- Text embedding models:
  - e5-mistral-7b-instruct (Microsoft Research) – also consider the large one, this is one mainly suited for English
  - Jina AI – English only (or German), 8k context length
  - BGE-M3 (Beijing Academy of Artificial Intelligence), multilingual, 8k context length
  - SBERT
- Object detection
  - OWL-ViT (Google): zero shot object detect. Can potentially be very useful for equiptment detection in images (cf. associated paper)
- Image to text:
  - UForm
  - Llava (Microsoft), should do OCR better
- Image embedding models:
  - UForm
  - CLIP: suited for both text and images, also consider multilingual variants if required.
- LLMs:
  - Mixtral-7B (Mistral AI): pure form requires around 90Gb of VRAM. Lower quantisation models are available.
  - Mistral-7B (Mistral AI), should run in a Colab. Lower quantisation models are available
  - Phi2 (Microsoft): mainly English, lower quantisations exist
  - Llama2 (Meta): has different versions, the smalles one should run on a colab, but lower quantisation models are available.
  - …
- BERTopic: clustering/grouping/topic detection, with multimodal and temporal capacities. Very flexible and modular.

Computational resources:

- The free Google Colab tier up to 15GB of VRAM may be enough to perform your calculations for many of the models (cf. practical sessions).
- The free Kaggle Tier allows up to 15GB of VRAM and 30Hrs of GPU per week
- You can use your own powerful desktop or laptop, of course, but keep in mind that most applications are optimised for Linux distributions.
- The department has some GPU resources (4 x 40Gb GPUs on an Ubuntu server, accessible via SSH) for some models that cannot be compressed.
- The HPC cluster has even more resources and may be employed as a last resort. The learning curve for this last choice is a little steeper.