Abstract: Measuring distances of related data content across multi-domain data, specifically face images and semi-structured text, Y1

Image processing of faces and topic modeling of text are both widely used forms of processing employed in machine learning applications. Facial features can be used to identify faces that look similar or in an ideal case, a specific person, depending on the quality of the image. Neural networks and support vector machines have become a widely used processing tool for feature extraction of images for this purpose. Topic modeling of text articles is employed across a variety of applications to improve content search from news or research articles, and is also used in a variety of natural language processing applications such as chat bots. The purpose of my research is to identify and validate candidate methods for linking textual articles and images they contain for assessing previously unseen images and/or articles. This requires that I transform both target data types (images and articles) into their respective n-dimensional spaces. I will use both convolutional neural nets (CNNs) and self organizing maps (SOM) to help me encode the different instances (image, article) and store them for indexing. Additionally, I will train another machine learning algorithm, yet to be identified, to classify new articles and images based on the trained article and image pairs. An additional aspect of my research is to explore whether or not there is a common n-dimensional functional space in which both vectorized images and vectorized text can exist together that has the properties suitable for comparing distances between the two. This would permit constructing a commons space for linking the two for measurement and comparison (e.g. an article and an image). The data sources will be news articles whose topics have been curated by the Gdelt data set (currently a 3 TB data set). Both natural language and image processing are computationally intensive, and so I am requesting an allocation on Bridges that permits use of multiple computing nodes and access to GPUs. Both the size of the data set and the computing requirements for the neural networks are beyond the capability of the systems I have at GMU.