Heterogeneous Data Fusion with Deep Architectures

Most Data Mining techniques are taught with the assumption that the data being processed is **homogeneous**, having uniform characteristics and traits. However, most interesting Business Intelligence problems in the real world require analyzing **heterogeneous** data, data consisting of various **modalities** (e.g., text, image, and structured data). For example, in the domain of **social tagging**, the goal is to find what kinds of labels a person will assign to an unknown entity, given a set of heterogeneous data that characterize the entity; for example, assigning tags to restaurants given both pictures and text descriptions. Algorithms that are robust to different data modalities could allow for businesses to find more general, robust patterns in data, and give rise to the notion of **entity discovery and recognition,** the ability to recognize a general entity no matter the data characterizing the entity.

In this project, we want to tackle the problem **of entity discovery and recognition** across various data modalities using the concepts outlined in the **Deep Learning** lecture material. First, we will take two real world datasets: one is a text corpus of stories from the BBC and the other is a repository of images from Flickr. Then, using the algorithm described in *Heterogeneous Network Embedding via Deep Architectures* by Chang et. al (2015), you will be asked to implement a **Deep Learning Architecture** that can guess the tags a user might apply to an image, while giving knowledge workers operating on the data insight into how the Flickr images relate to words or concepts (e.g., is this image related to birds, which might be related to flight, etc).

Goal: To implement a social tag discovery algorithm where items determined “similar” should share a tag or label, with the similarity metric being a normalized dot-product between two *simultaneously learned vector space embeddings* for each item in the data source.

Project Requirements

To implement this algorithm, first read *Heterogeneous Data Fusion via Deep Architectures* by Chang et. al (http://dx.doi.org/10.1145/2783258.2783296).

You will be provided with the following materials in advance (on Moodle):

- A data folder (~300MB) containing the bbc text corpus and a matrix file for 127-by-127 grey-scale samples of the flickr dataset (dealing with birds and planes).
- A project file `datafusion.py` with the project implementation results.

Project Details:

- Implement the deep learning architecture described in Section 4 of the publication by following the text instructions in `datafusion.py`. Look for the tag `# YOUR CODE GOES HERE` and fill in the missing parts to complete the code. **IMPORTANT**: make sure you have version 6.0 of TensorFlow installed on your system. The TensorFlow git repo has instructions for installing this version.

Submission Instructions:

- Make your changes directly to the `datafusion.py` file and submit this file on Moodle. Make the filename `unityid_datafusion.py`.