Deep Learning - Project proposal

Group: Alexander Krohannon, Sharmila Selvaraj, Gayathri Panangipalli

**Introduction**:

Publishing articles in scientific journals is a necessary and daunting part of researchers’ jobs across the globe. However not all research is created equal, and each journal, in addition to subject specifications, has specific standards as to what quality of research they are willing to publish. A commonly used measure to quantify research quality is called impact factor of an article. This is measured in terms of the number of citations during the first six months of its publication. And various journals specialize in publishing works that will have an impact factor within a specific range: Biological Reviews 4.966, Nature Methods: 21.18, Molecular Systems Biology 7.040, BioScience 2.709, Journal of Mathematical Biology 0.903.

**Problem Statement:**

Whether its developing novel methods, discovering a previously unknown association, optimizing an existing process, or anything else in the wide world of scientific inquiry, performing research is difficult. But figuring out which journals are likely to accept and publish a paper shouldn’t be. To this end, we aim to create a deep learning model capable of taking an article abstract as input and predicting the impact factor the article will have. This citation metric could then serve to help tailor the researcher’s search to only journals that are likely to publish articles with similar metric.

**Datasets used:**

Dataset for this study will be taken from National Library of Medicine (NLM) baseline set of MEDLINE/PubMed citation records. The PubMed baseline repository is updated regularly with the last update being on February 03 2020. It has a collection of citation records pertaining to biomedical literature from varied sources such as journals and online books. The data is stored in a XML format containing publication and citation records for each article along with abstract and author information.

**Method overview:**

1. Data download and processing: Data will be downloaded from the NLM website in a XML format and parsed using python packages. Required columns such as citation records, abstract will be extracted.
2. Citation metric: Citation metric will be used as a label for the model. NetworkX – python package will be used for getting the number of citations per article by calculating the number of incoming edges for each article. Each node in the network represents an article and directed edges represent the number of citations for that article. To reduce complexity of the network, an edge for a node will be considered if the article citing it, is published within 6 months of it’s respective cited article.
3. The model: Since the data given to the model will be that of research articles, an LSTM-RNN architecture would be implemented as it considers the sequence context which is a highly desired feature while looking at long range semantic texts. In order to account for the novelty of concepts as well as the quality of work, journals usually look for logical flow of information in an abstract. This deep-learning architecture can capture sequential information and hence the problem of misinterpretation of abstracts meaning can be eliminated. This will be done using Keras framework.
4. Model implementation: For input to the model, data downloaded will be processed using NLTK packages. An 70:10:20 split ratio for training, validation, and test data will be performed. Citation metric and the abstract would be fed into the model. Based on the learning from train data, parameter tuning will be done to obtain better performance.
5. Model validation: A list of articles independent of the above dataset published after February 2020 with the needed information, will be taken from PubMed or Web of Science and used as a validation set.

**Progress:**

Methods for predicting the articles impact factor have been decided. Data for this study has been downloaded and is yet to be processed.

**Plans ahead:**

1. Parsing XML file to get the desired columns.
2. Pre-process the text from string to numerical format using processes such as tokenizing and padding (if needed).
3. Calculation of citation metric using networkX package.
4. Model implementation and parameter tuning based on initial results.
5. Hyperparamter tuning based upon validation results
6. Evaluation of model performance.