Nisarg Mistry (nmistry2@illinois.edu) / Sandeep Kumar (kumar64@illinois.edu)

CS 598 DEEP LEARNING FOR HEALTHCARE - PROJECT PROPOSAL

Paper 111: Real-world Patient Trajectory Prediction from Clinical Notes Using Artificial Neural Networks and UMLS-Based Extraction of Concepts

1. **Citation to the original paper**

   Link: https://link.springer.com/article/10.1007/s41666-021-00100-z

   Citation: Zaghir, J., Rodrigues-Jr, J.F., Goeuriot, L. et al. Real-world Patient Trajectory Prediction from Clinical Notes Using Artificial Neural Networks and UMLS-Based Extraction of Concepts. J Healthc Inform Res 5, 474–496 (2021). https://doi.org/10.1007/s41666-021-00100-z

2. **What is the general problem this work is trying to do? We are not asking for the specific approach, that's requested below. An example of a general problem is 'mortality prediction.' An example of a specific approach is 'using recurrent neural network and attention mechanism.' Do not copy the description in the paper – use your own rewording.**
   The work in this paper tries to perform computer-aided medical prognosis using textual clinical notes extracted from EHR (Electronic Health Records) from the MIMIC-III dataset. The data is passed through neural networks to predict the probable medical problems of the patient in the future like clinical conditions, mortality, and readmissions. The paper tries to extract information from unstructured text found in clinical notes and tries to predict the most probable medical conditions using deep neural networks.

3. **What is the new specific approach being taken in this work, and what is interesting or innovative about it, in your opinion?**
   The previous papers used word embeddings and raw sets of concepts extracted from the medical clinical notes for performing predictions. However, this approach tries to use a refined set of concepts called UMLS (Unified Medical Language System) which applies similarity-based threshold filtering and a list of acceptable concept types and tries to reduce the number of probable diagnosis codes. The same deep learning models are used for multiple problem statements to predict the mortality rates, readmission, and diagnosis codes.

4. **What are the specific hypotheses from the paper that you plan to verify in your reproduction study?**
   As part of our work, we plan to verify whether the usage of a unique concept of UMLS (Unified Medical Language System) performs better than the standard coding systems like ICD codes. Also, we try to verify the results achieved by the paper for various types of tasks like predicting the diagnosis codes, mortality, and readmission probabilities.

5. **What are the additional ablations you plan to do, and why are they interesting? (Open)**
   Along with the model architectures in the paper, we plan to modify the neural network models to see if it is possible to achieve results better than the current models.

6. **State how you are assured that you have access to the appropriate data.**

The dataset being used in this paper is the clinical notes over the MIMIC-III dataset. We have requested access to the dataset as it is a freely available dataset.

7. **Discuss the computational feasibility of your proposed work – make an argument that the reproduction will be feasible.**
   The model in the original paper is trained using hardware NVIDIA Quadro P6000 GPU with around 5000 epochs. We would be using Google Collab to train using GPU.

8. **State whether you will re-use existing code (and provide a link to that code base) or whether you will implement yourself.**
   For data preprocessing we would be using the code available: https://github.com/JamilProg/patient_trajectory_prediction. We would try to build on our deep learning model along with the already built network architectures.

Paper 137: Automated ICD-9-CM medical coding of diabetic patient's clinical reports

1. **Citation to the original paper**
   URL: https://dl.acm.org/doi/10.1145/3279996.3280019

   Vitor Pereira, Sérgio Matos, and José Luís Oliveira. 2018. Automated ICD-9-CM medical coding of diabetic patient's clinical reports. In Proceedings of the First International Conference on Data Science, E-learning and Information Systems (DATA '18). Association for Computing Machinery, New York, NY, USA, Article 23, 1–6.
   DOI:https://doi.org/10.1145/3279996.3280019

2. **What is the general problem this work is trying to do? We are not asking for the specific approach, that's requested below. An example of a general problem is 'mortality prediction.' An example of a specific approach is 'using recurrent neural network and attention mechanism.' Do not copy the description in the paper – use your own rewording.**
   In this paper, the authors are proposing an approach to building a system that can help automate the assignment of ICD-9-CM codes to clinical records. Assignment of ICD-9-CM code to medical report is very important in the medical industry, done by dedicated professionals and costs approx. 25B USD. ICT-9-CM helps in coding treatments or procedures carried out on patients and correct coding helps all players in the medical industry to standardize data, reduce fraudulent transactions and mitigate financial loss risk.

3. **What is the new specific approach being taken in this work, and what is interesting or innovative about it, in your opinion?**
   The authors are proposing two CNN architectures - single dimension and three parallel convolution layers. What is interesting is the approach to feed reports' category also as an input and the outcome of it - the baseline CNN model F1 score decreased by 9% and models without any dense layer did not change significantly. This experiment underlines that additional input

data may not help improve the performance of a CNN, especially when there is no clear way for the CNN model to utilize the input data and learn any differentiating feature from it.

4. **What are the specific hypotheses from the paper that you plan to verify in your reproduction study?**
   I would like to validate the hypothesis that category augmented information does not produce a better result.

5. **What are the additional ablations you plan to do, and why are they interesting?**
   The paper uses word2vec for the tokenization of medical records. We would like to try Med2Vec as well to see the behavioral difference. Assuming that the medical records should be rich with medical terminologies, Med2Vec, specially designed for the medical domain, may yield better word embedding as compared to general Word2Vec.

6. **State how you are assured that you have access to the appropriate data.**
   Paper uses publicly available MIMIC III data - NOTEEVENTS and DIAGNOSIS_ICD which is available on physionet.org and we have secured access to the MIMIC III dataset on physionet.org.

7. **Discuss the computational feasibility of your proposed work – make an argument that the reproduction will be feasible.**
   Based on the volume of preprocessed data statistics, it seems feasible to reproduce the proposed work on a personal laptop - our first preference. However, in case of difficulties - memory requirements, too long to train the model, we propose the following alternatives - 1. Training with a reduced dataset - 2nd preference, 2. Training on public cloud infrastructure with GPU - worst-case scenario

8. **State whether you will re-use existing code (and provide a link to that code base) or whether you will implement yourself.**
   We do not have access to the existing code, and we plan to implement the CNN models ourselves.

Paper 164: DistCare: Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis

1. **Citation to the original paper**
   URL: https://dl.acm.org/doi/10.1145/3442381.3449855

   Liantao Ma, Xinyu Ma, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Chaohe Zhang, Wenjie Ruan, Yasha Wang, Wen Tang, and Jiangtao Wang. 2021. Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis. In <i>Proceedings of the Web Conference

2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 3558–3568. DOI:https://doi.org/10.1145/3442381.3449855

2. **What is the general problem this work is trying to do? We are not asking for the specific approach, that's requested below. An example of a general problem is 'mortality prediction.' An example of a specific approach is 'using recurrent neural network and attention mechanism.' Do not copy the description in the paper – use your own rewording.**
   In this paper, the authors are highlighting a common issue related to the lack of sufficient data to help the prognosis of emerging infectious diseases, in the early days of pandemics such as CoVid 19.

3. **What is the new specific approach being taken in this work, and what is interesting or innovative about it, in your opinion?**
   The authors are proposing a distilled transfer learning framework, DistCare, which leverages publicly available EHR datasets to augment the prognosis of emerging infectious diseases. The model learns to embed the CoVid-19 related features on massive existing EHR data, and the transferred parameters are further trained to imitate the teacher's model representation based on distillation.

4. **What are the specific hypotheses from the paper that you plan to verify in your reproduction study?**
   In the reproduction, we would like to verify the effectiveness of transfer learning, proposed in the paper.

5. **What are the additional ablations you plan to do, and why are they interesting?**
   It would be worth checking the performance of the transfer model utilizing a public dataset for similar diseases e.g., infectious diseases with similar mortality rates vs. any infectious disease.

6. **State how you are assured that you have access to the appropriate data.**
   Paper uses PhysioNet dataset and CoVID-19 dataset for HM Hospital. We have already secured access to the PyhsioNet dataset and are planning to reach out to HM hospital for the CoVid-19 dataset.

7. **Discuss the computational feasibility of your proposed work – make an argument that the reproduction will be feasible.**
   Experiment environment utilized in the paper - CPU: Intel Xeon E5-2630, 256GB RAM, and GPU: NVIDIA TitanX. We plan to use Google CoLab, with an option for additional CPU and GPU units.

8. **State whether you will re-use existing code (and provide a link to that code base) or whether you will implement yourself.**
   Given the complexity of the implementation, we plan to use the existing code base https://github.com/Accountable-Machine-Intelligence/DistCare as a reference and we try to implement the proposed model on our own.