Reproducibility Report for “Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives”

Jeremy Bao and Luke Freitag
{jbao8, lukegf2}@illinois.edu

Group ID: 50
Paper ID: 163

1 Cite the original paper.

We will be working on replicating and extending the experiments performed in the paper “Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives” [1]. A citation for that paper can be found at the end of this document.

2 State the general problem the paper aims to solve. Do not use the same language as the paper.

According to this article, the creation of automated methods capable of determining the characteristics of patients based on natural language clinical texts is an important task, as these texts contain a lot of information and detail not included within the various medical codes stored in those patients' EHRs, but extracting these characteristics manually is time-consuming and error prone. Additionally, it is important that these methods can, when claiming that a patient has a certain characteristic, explain why they made their decision. This is because medical personnel may not trust the output of these methods unless they can output an explanation which aligns with scientific knowledge, strange explanations may help uncover problems with the data and faulty assumptions made during research, and governments such as the European Union have been considering making laws to prohibit the usage of machine learning algorithms which cannot explain themselves properly.

3 Describe the new and specific approach taken by the paper. Discuss why it is interesting or innovative.

Previous methods for extracting patient characteristics from clinical notes can be divided into two categories. In both types of methods, medical experts first have to create dictionaries of phrases relevant for determining if a patient has a certain characteristic, and the occurrences of those phrases in each document being investigated must be counted. Many libraries capable of finding phrases from these dictionaries within documents have already been created, allowing researchers to deal with misspellings, synonyms, and ambiguous terms without having to think of them independently.

Methods from the first category involve using machine learning models to make predictions about the patient described in each document, based on the counts for the relevant terms within it. Methods from the second category involve creating lists of rules based on the presence and absence of these relevant terms to determine if the patient described in each clinical note has a characteristic of interest.

Both of these approaches can be labor intensive (as phrase dictionaries may have to be combined in both cases, as it may take effort to decide which phrases should be included and excluded in both cases, and as rules must be carefully crafted in the second), create models which are specific to a single task and often do not generalize well between hospitals, and are not very useful in real world conditions.

This paper's authors decided to leverage advances in deep learning techniques to create convolutional neural networks (CNN) capable of reading discharge summaries, learning which phrases in them correspond to patients being likely to have or not have ten phenotypes, and deciding which of those conditions a patient has. The phenotypes investigated were advanced or metastatic cancer, advanced heart disease, advanced lung disease, chronic neurologic dystrophies, chronic pain, alcohol abuse, substance abuse, obesity, psychiatric disorders, and depression.

The CNNs described in this paper were fairly simple. Embedding vectors for each word were creating using Word2vec on all discharge summaries within the MIMIC-III dataset. The 1,610 discharge summaries to be used in training, validation, and testing were were apparently padded to an equal length, then converted to 2-dimensional grids with a height equal to the embedding vector length and a width equal to the document length. Multiple convolutional layers with different kernel sizes, each of which included multiple kernels, were then applied to these grids. Each kernel would read in the embedding vectors of 1, 2, 3, 4, or 5 adjacent words at a time, and output a single value corresponding that section of text. A max pooling operation would be applied to the output of each kernel, so that a vector containing the maximum output value from sliding each kernel of each size over the document would be created. Finally, that vector would be passed into a fairly small fully-connected layer outputting the probability of the patient having the phenotype in question. A separate CNN would be used for predicting if the patient had each of the phenotypes being investigated (so we would have one CNN for predicting advanced or metastatic cancer, another for predicting advanced heart disease, and so on).

This approach is interesting and innovative, as it involves the application of neural networks into an area previously dominated by traditional machine learning techniques, as it requires less human input and domain knowledge than previous methods, as it can be generalized to different tasks and between hospitals more easily, and as it achieved much better performance than previous approaches for detecting most patient phenotypes. Also, while neural networks are often thought of as being obtuse black boxes, the creators of this paper implemented a way to figure out which words and phrases had the biggest impacts on the decisions their CNNs made.

4 Identify the specific hypotheses you plan to verify in your reproduction study.

We plan to verify that the CNN designs described in the paper can achieve higher performance for predicting whether patients have or do not have the ten phenotypes described above, compared to the other methods they discussed. We will construct the CNNs from the paper, train them, and compare the results they achieve with the results of the other models they described (including a logistic regression model operating on a vector of word counts, a logistic regression model operating on a vector of n-gram counts, for n-grams up to size 5, and random forest, naive Bayes, and logistic regression models operating on the results of using the cTAKES library to count either all medical terms or medical terms relevant to the phenotype in question within each document). We will implement, train, and test some or all (depending on time, capability to understand them, and data accessibility) of these other models as well, so that we can compare results. We will also investigate if considering longer phrases (by including filters with larger sizes) increases performance for the CNNs, as it did for the creators of this study.

The creators of this study also assessed the interpretability of their model by extracting the words and phrases which contributed most to their CNN deciding if a patient had or did not have some phenotype and comparing them to the words and phrases domain experts described as being related to the phenotype in question. They then did the same for the random forest models which represented documents as counts of phrases deemed relevant to the phenotype in question generated by cTAKES, and found that the terms which contributed most to the decisions of the CNNs were closer to the ones specified by the experts than the ones which contributed to the decisions of the random forest models.

However, it may be hard for us to replicate this investigation of interpretability. One reason for this is that we do not understand the mathematics involved in the process of determining the phrases which most influenced the decisions of a CNN, and the explanations given were not extensive. Another reason is that that we do not have access to the lists of terms deemed by medical experts to be most relevant for deciding if patients had the various phenotypes or not, and lack the time and domain knowledge needed to create them. Therefore, even if we managed to figure out which phrases were most influential, we would be unable to tell if the CNNs' decision making processes were in line with current medical knowledge. Because of this, we might investigate this further if we have time available.

5 Outline any additional ablations you plan to do and explain why they are interesting.

There are various ablations we could perform to investigate the importance of different components of the neural networks described in this paper. One possible ablation would be to represent words as one-hot vectors instead of embedding vectors (produced by running Word2Vec on all discharge summaries within MIMIC-III in the original study). This would be interesting, because we learned that embedding vectors can greatly improve the performance of machine learning models in class, by having similar words represented in a similar way, instead of having all words be considered equally different. For instance, in a simple one-hot encoding system, the word "dog" might be represented as $< 1, 0, 0 >$, "cat" might be represented as $< 0, 1, 0 >$, and "fish" might be represented as $< 0, 0, 1 >$, so that all of these things have the same distance from each other. However, cats and dogs are both mammals, and are more similar to each other than they are to fish. Representing these words using embedding vectors would probably make the representations for "cat" and "dog" similar to each other and different from the representation for "fish," which reflects reality better. We would accomplish this ablation by omiting the step in which words are converted into embeddings, and do everything else in the same way.

Another possible ablation would be to omit the max pooling step applied to the output of the parallel convolutional layers. This would be interesting, as we learned that pooling is often useful for reducing the effects of translation in CNNs. Seeing how performance is affected by the removal of the pooling layer would help us see how much of an effect it has. However, we have heard that pooling layers can increase performance by reducing the size of the input fed into the next layer, and are not sure if carrying this ablation out would still allow us to complete our training in a reasonable length of time.

6 Explain how you have access to the necessary data.

The data used in this study consists of 1,610 discharge summaries from MIMIC-III. We have access to MIMIC-III via PhysioNet, and can probably find those discharge summaries by joining the table found in data/annotations.csv in the GitHub repository provided by the authors of this paper with the table found in NOTEEVENTS.csv.gz from MIMIC-III.

The provided repository also provides a link to the Word2vec embeddings used by the authors of the study. However, this link is dead. We probably lack the resources to generate these ourselves, as we do not have enough computational resources to analyze all of the hundreds of thousands of documents the authors used. We may try to contact the authors of the study to ask them for the file containing the embeddings. If we cannot get access to that file, we may instead run Word2vec on the 1,610 documents used in this study, or something like that.

7 Discuss the computational feasibility of your proposed work.

Replicating this study should be fairly feasible, as the dataset which will be used is not very large and contains only 1,610 documents, and the documents do not seem very large. The models we will construct are also just a mix of traditional machine learning models and a fairly small CNN.

8 Specify if you will be re-using existing code and provide a link to it, or if you will implement the code yourself.

The researchers provided a link to the GitHub repository containing their code, which can be found at https://github.com/sebastianGehrmann /phenotyping. However, only parts of their code (for preprocessing data and carrying out the benchmark machine learning models) is in Python. The parts of their code responsible for their CNN and perhaps for other things were written in Lua, and neither of us have experience with that language.

We may look through the code provided by this study's authors for inspiration, especially when creating code for preprocessing the data, implementing the various benchmark models, and figuring out the details of the architecture of the CNNs to be used. However, we will write our code ourselves, especially the parts needed to create, train, and test the CNNs for predicting the presence and absence of the discussed phenotypes (this will be done in Python, using PyTorch).

References

[1] Sebastian Gehrmann et al. "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives". In: *PloS one* 13.2 (2018), e0192360.