Embedding-Augmented Deep CNNs for PudMed Journal Recommendation

Calvin Chi

Introduction

- It is costly to submit a scientific paper to the wrong journal
- Journals are often characterized by their research topic, significance of results published, or even writing style
- Editorial offices can sometimes determine the "fit" of a submitted paper based on the abstract or introduction
- The goal of this project is to evaluate the predictive ability of abstracts for journals in PubMed
- The objectives of this project are to answer the questions
  1. Do abstracts contain information that can be used to infer the journals they are accepted by?
  2. How important are research topic and finding significance, as could be inferred from abstracts, for journal prediction?

Figure 1. Word cloud of *Nature* abstracts

Figure 2. Word cloud of *Science* abstracts

Data

- A total of 415,381 PubMed abstracts, all annotated with (1) research topic (from search result), (2) impact factor, (3) journal published in
- Restrict to topics in Mendelian diseases, statistical genetics, population genetics, bioinformatics, omics technologies, genome structure, epigenetics, and developmental genetics
- Restrict to articles published within 10 years of 2018
- Remove journals with less than 40 abstracts in dataset to result in dataset with 1,548 unique journals

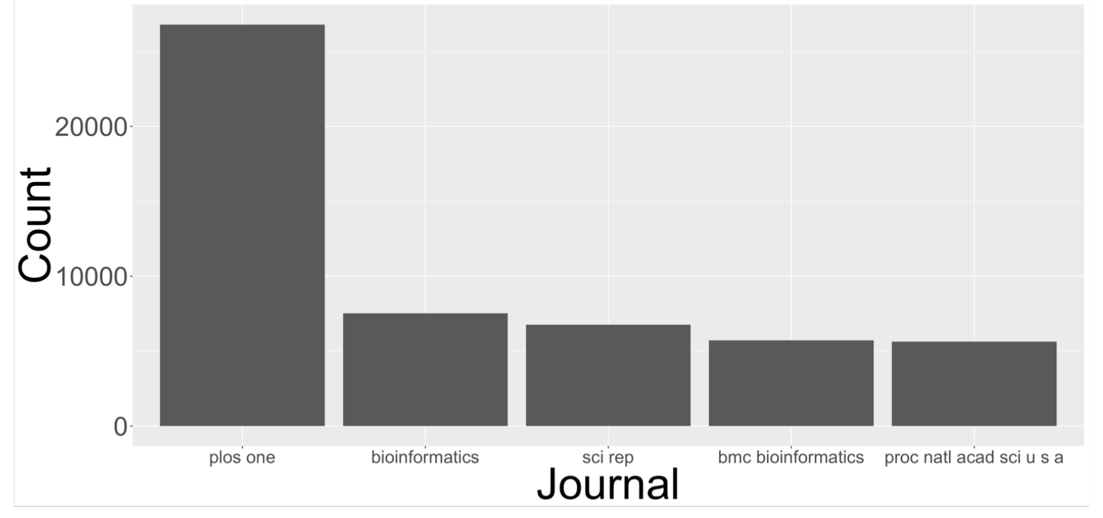

Figure 3. Count of top five common journals

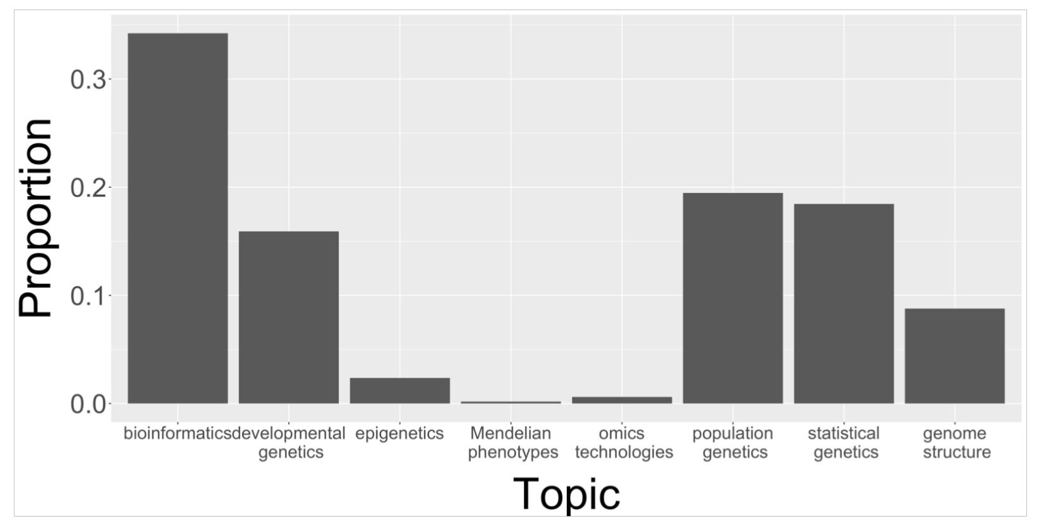

Figure 4. Distribution of research topics

Methods

- Input word embeddings of dimension 200, pre-trained on 700,000 PubMed articles with word2vec of window size of five
- Impact factors relabeled as 5 bins based on four quartiles
- Adam optimizer with categorical cross entropy loss for two epochs
- Convolutional neural network (CNN) architectures evaluated
  - Standard CNN architecture from Yoon Kim in 2014
  - Multitask learning of research topic, impact factor, and journal with CNN
  - Embedding-augmented CNN
- Basic CNN architecture
  1. Fixed embedding layer of pre-trained weights
  2. First convolution layer: 1D convolution of 128 filters of size 5 with ReLU, batch normalization, max pooling of window size 5
  3. Second convolution layer: 1D convolution of 128 filters of size 5 with ReLU, batch normalization, max pooling of window size 5
  4. Dense layer of output 128 units with ReLU
  5. Dense output layer with softmax
- Obtain embeddings from CNNs trained to predict research topic and impact factor respectively
- Embedding-augmented CNN: concatenate topic and impact factor embeddings with output of convolutional layers, followed by fully-connected layers for journal prediction

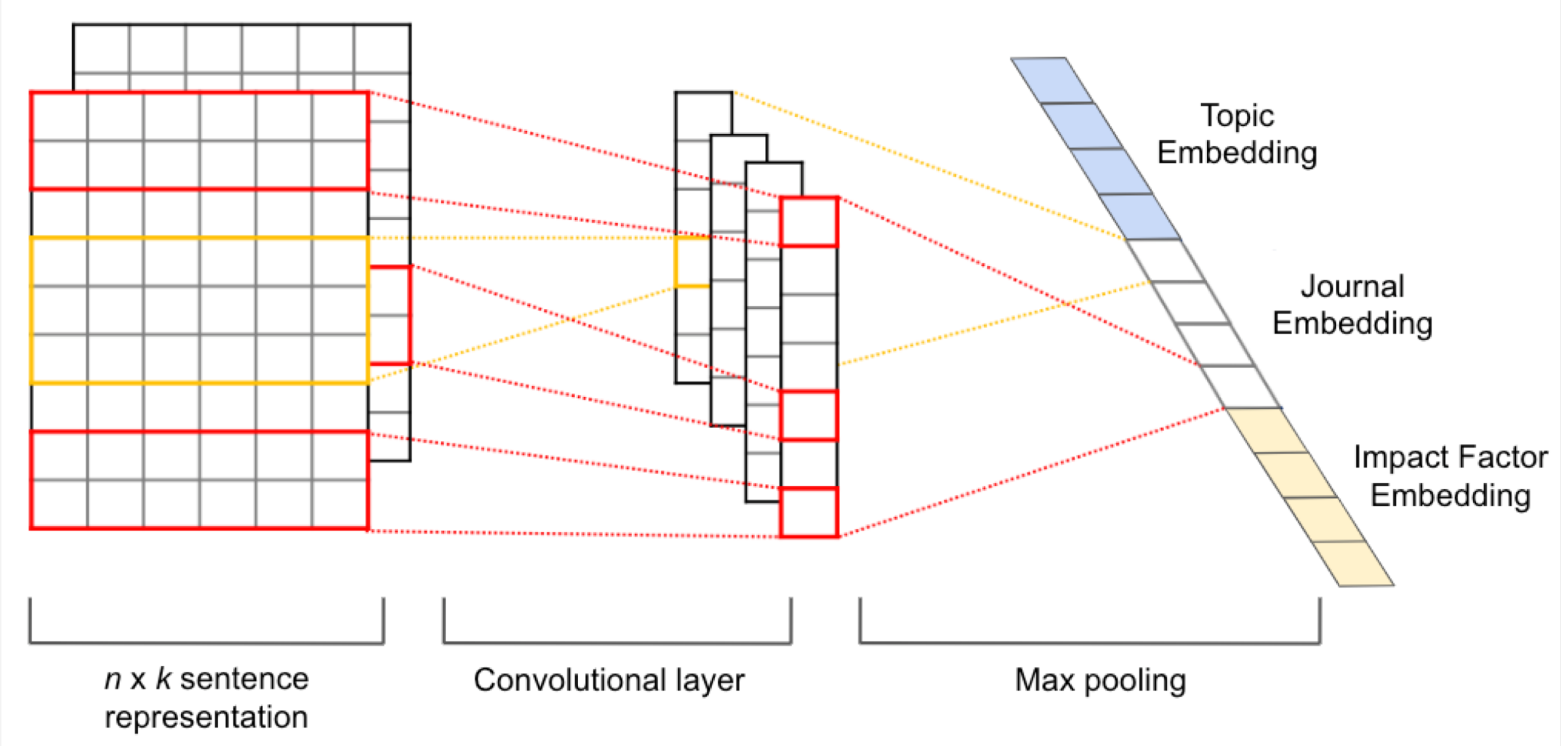

Figure 5. Graphic illustration of augmenting output of convolutional layers with topic and impact factor embeddings

- Let $H_k^{(i)}$ be set of top k journal predictions and $y^{(i)}$ be journal label for abstract i, the coverage accuracy at k is defined as

$$\mathcal{A}(y, H_K) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y^{(i)} \in H_k^{(i)})$$

- AUC: area under curve of coverage accuracy vs k

Results

- The embedding-augmented CNN achieves best performance uniformly across the metrics of accuracy, AUC, and k to achieve coverage accuracy of at least 90%

Model	Accuracy	AUC	K for 90% coverage accuracy
Baseline CNN	21.6%	0.975	80
Multitask CNN	19.1%	0.971	100
Embedding-augmented CNN	23.7%	0.978	60
Majority Class	6.6%	NA	NA

Table 1. Table of performance on 10% test dataset for different architectures

- Trained embeddings for research topic and impact factor show separation between different categories

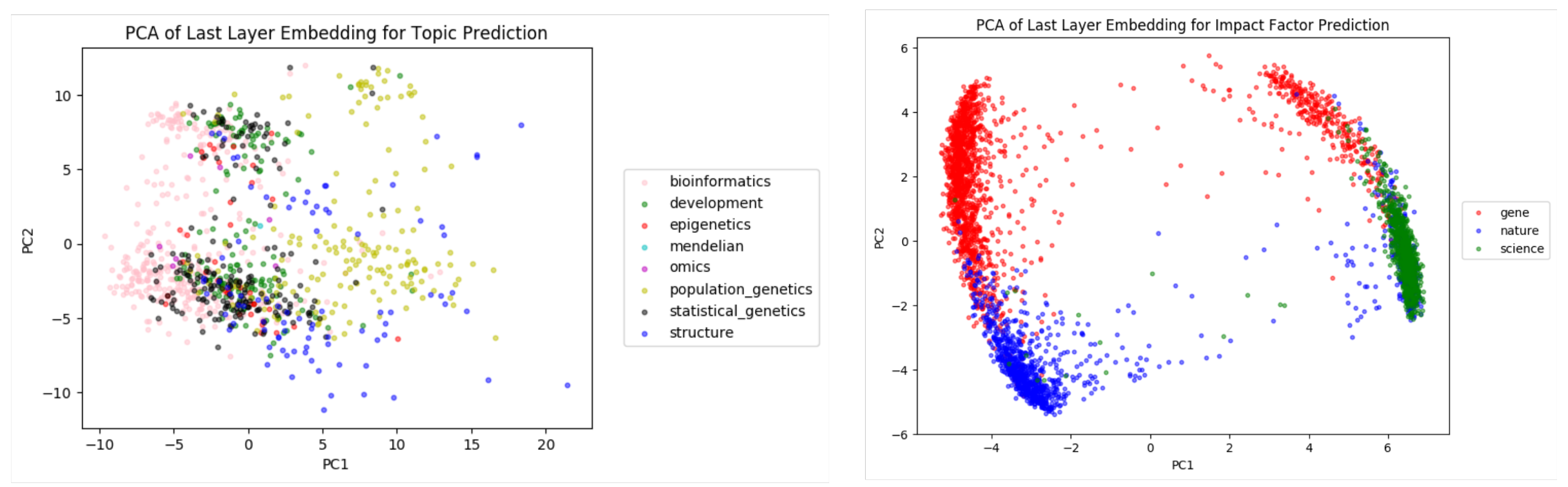

Figure 6. PCA of trained embedding for (a) research topic and (b) impact factor.

- All CNN models are effective at predicting the right journal within the first few predictions

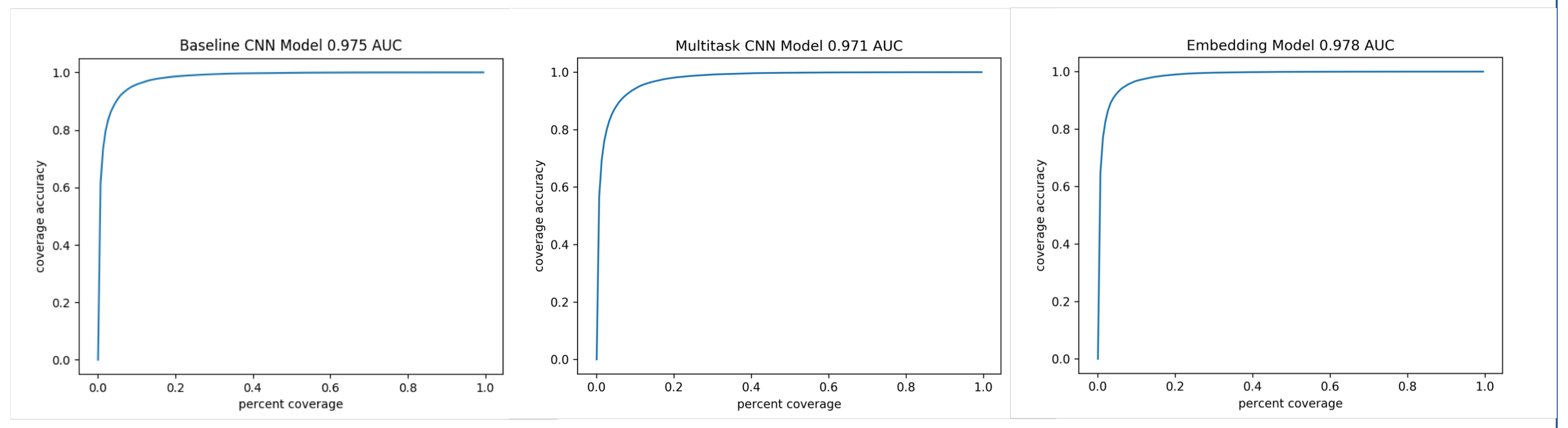

Figure 7. coverage accuracy vs percent coverage for (a) baseline CNN (b) multitask CNN and (c) embedding-augmented models. Percent coverage defined as k / number of distinct journals

Conclusions

- Information on research topic and significance improve journal prediction from abstract text
- Research topic and results significance are not the only information relevant for journal prediction since multitask CNN performed the worst