National Human Genome Research Institute

The Forefront of Genomics

Applications of Deep Learning in Medical Genetics

Simon Liu[1*], Suzanna Ledgister Hanchard[1*], Michelle C. Dwyer[1*], Ping Hu[1], Cedrik Tekendo-Ngongang[1], Rebekah L. Waikel[1], Dat Duong[1], Benjamin D. Solomon[1]

[1] Medical Genomics Unit, National Human Genome Research Institute, Bethesda, MD, *co-first authors

Introduction

- Artificial intelligence (AI) is increasingly studied and applied in a variety of biomedical contexts. Among AI, **deep learning (DL) has shown strong potential in recent breakthroughs**.
- Unlike other machine learning (ML) subbranches, **DL does not require domain experts to generate features**.
- Medical genetic conditions are **complex esoteric, rare, and often less-well studied**. However, with the advancement of computational technology, we expect DL to become an **increasingly important tool in analyzing medical genetics datasets**.
- We present a DL classifier to **automatically identify DL publications in medical genetics with minimal manual curation,** which can be further adapted to other text-based analyses.

Data Curation

- Of the 14,002 article PMIDs collected in our initial scoping analysis, 306 were excluded due to **unavailability** (295), **corrections** (6), and **duplicates from preprint servers** (5).
- Abstracts for the remaining **13,696 articles** were downloaded using the eDirect CLI Utility.[1]

Methods

- Two models used for classification: random forest (**RF**) and Bidirectional Encoder Representations from Transformers (**BERT, Fig. 1**)
- We are primarily interested in the binary classification of **category 1 (DL in medical genetics)** versus the other categories.
- Data split 8-1-1 for 9-fold cross validation for both models, yielding **9 RF models**, **9 BERT models**, and **15 string-search labeling functions**.
- Snorkel[2] used to aggregate classifier and labeling function predictions

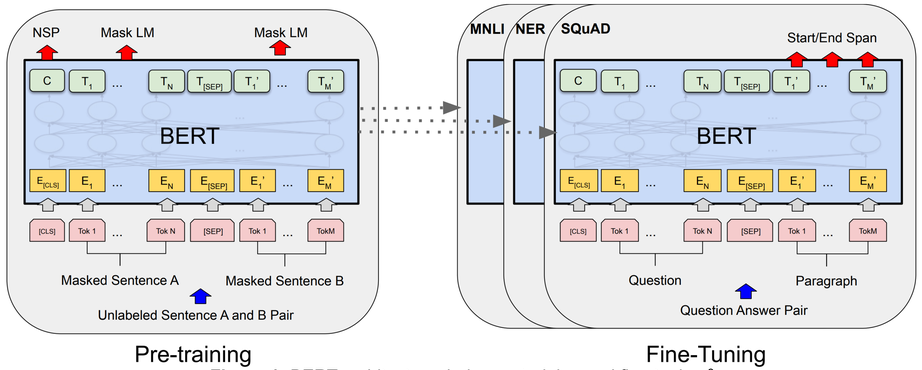

Figure 1: BERT architecture during pretraining and fine tuning.[3]

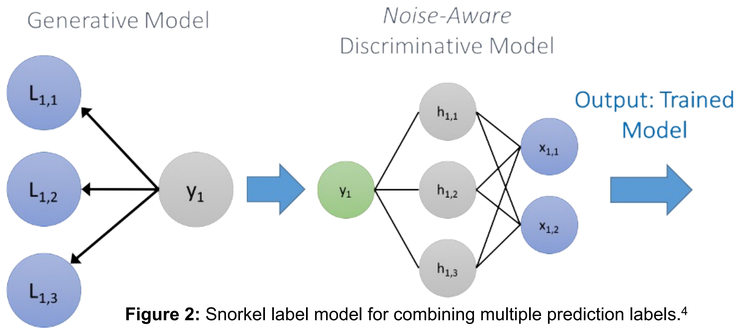

Figure 2: Snorkel label model for combining multiple prediction labels.[4]

Results

Snorkel Model	TPR	FPR	FNR	TNR
LFs only	0.143	0.378	0.571	0.469
RFs only	0.214	0.000	0.786	1.000
BERTs only	**1.000**	**0.055**	**0.000**	**0.945**
LFs + RFs	0.143	0.378	0.643	0.469
LFs + BERTs	0.143	0.378	0.857	0.622
RFs + BERTs	**1.000**	**0.055**	**0.000**	**0.945**
LFs + RFs + BERTs	0.143	0.378	0.857	0.622

Table 1: labeling functions (**LFs**), random forests (**RFs**), true positive rate (**TPR**), false positive rate (**FPR**), false negative rate (**FNR**), true negative rate (**TNR**)

Discussion

- With a **100% TPR** and **5.5% FPR**, our approach can be very useful in automatically reducing a large set of papers that need to be manually assessed for relevance.
- Manual review of the 76 false positives did identify one article that was **initially incorrectly categorized** as category 4 and should have been category 1.
- Future applications include generating a "likelihood of diagnosis" from Undiagnosed Diseases Program notes, an ongoing project analyzing diagnostic trends.

References

1. Kans, J. (2013, April 23). *Entrez Direct: E-utilities on the Unix Command Line*. National Center for Biotechnology Information. Retrieved April 26, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK179288/
2. Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017, November 28).. arXiv.org. Retrieved April 26, 2022, from https://arxiv.org/abs/1711.10160 *Snorkel: Rapid Training Data Creation with weak supervision*
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
4. Ghelani, S. (2019, June 2). *Snorkel-a weak supervision system*. Medium. Retrieved April 25, 2022, from https://towardsdatascience.com/snorkel-a-weak-supervision-system-a8943c9b639f

All relevant data and code can be found via the QR code or at https://github.com/simonliu99/classify-medical-genetics-abstracts.