Frequency Bias in MLM-trained BERT Embeddings for Medical Codes

T. Yu[1], T. Tuinstra[1], B. Hu[1], R. Rezai[1], T. Fortin[1], R. DiMaio[1], B. Vartian[2] and B.P. Tripp[1]

[1] Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada
[2] Department of Family Medicine, McMaster University, Hamilton, Canada

Abstract— **Transformers are deep networks that operate on loosely structured data such as natural language and electronic medical records. Transformers learn embedding vectors that represent discrete inputs (e.g. words; medical codes). Ideally, a transformer should learn similar embedding vectors for two codes with similar medical meanings, as this will help the network make similar inferences given either of these codes. Previous work has suggested that they do so, but this has not been analysed in detail, and work with transformers in other domains suggests that unwanted biases can occur. We trained a Bidirectional Encoder Representations from Transformers (BERT) network with clinical diagnostic codes and analyzed the learned embeddings. The analysis shows that the transformer can learn an undesirable frequency-related bias in embedding similarities, failing to reflect true similarity relationships between medical codes. This is especially true for codes that are infrequently used. It will be important to mitigate this issue in future applications of deep networks to electronic health records.**

Keywords— **Masked Language Modeling, Embeddings, BERT, Medical AI, Electronic Health Records**

I. Introduction

The last ten years have seen exponential growth in the volume of digital medical data [1] with the widespread adoption of Electronic Health Records (EHRs), giving rise to the possibility of deep network-based prediction of medical events with high accuracy [2].

EHRs are well suited [3] for processing by transformers [4], a recent kind of deep network that is extensively used in language-related applications. The input to a transformer is a sequence of elements from a pre-defined vocabulary, such as words and/or medical codes. Vocabularies typically contain tens of thousands of items. Transformers learn a unique vector, called an embedding, to represent each vocabulary item.

A transformer should learn similar embedding vectors for similar vocabulary items (e.g the words ”big” and ”enormous”). A previous application of transformers to medical records [5] suggested success in this respect. Specifically, they reported that among the ten most similar embeddings for each of the 87 most frequently used disease codes, 76% corresponded to clinically valid relationships. However, in that work clinical validity was not clearly defined, and this analysis was not performed for embeddings of less common codes. Meanwhile, in natural-language applications of transformers, the relative frequency with which words appear in a training corpus has been seen to result in biases in the embedding space [6, 7]. The embeddings of low and high-frequency words lie in different subregions after training, and due to differences in frequency, some semantically similar words are distant from each other in the embedding space [6, 7]. This phenomenon has potential implications for transformers in the medical domain, raising concerns for appropriate prediction of uncommon disease codes [8, 9]. Clinical applications of transformers could be even more susceptible to this issue than language applications, because medical training data is limited relative to internet-scale text corpuses. This is a concern due to safety implications of medical applications.

This paper analyses disease-code embeddings in a transformer trained on medical records. We show that frequency-related biases form systematically in the embedding space. We study several examples in more detail, and show that the transformer sometimes learns inappropriately similar embeddings for poorly related medical concepts, and non-similar embeddings for closely related medical concepts. Finally, we show that the transformer’s predictions are biased toward more common diseases.

II. Methods

A. Data

We used the Medical Information Mart for Intensive Care (MIMIC) dataset, a collection of de-identified EHRs collected from hospital patients [10]. We used the data in MIMIC-IV v2.0, which contains information from over 315,000 patients collected between 2008 and 2019. MIMIC-IV is freely available through PhysioNet [11] under a license and data use agreement.

We used the ICD-9 and ICD-10 disease codes in the *hosp* module of the MIMIC-IV dataset. To pre-process the data, we filtered out patients who did not have at least one entry in the *icd_diagnosis* table. For each of the remaining 190,121 patients, their hospital admissions were sorted by time and assigned an integer, starting from 1, to represent the order of hospital visits. Each unique ICD-9 and ICD-10 code in the dataset was assigned a token using a WordLevel tokenizer from the Hugging Face transformers library [12]. In total, there were 26,268 tokens, including the special tokens: `[PAD]`, `[CLS]`, `[SEP]`, `[MASK]`, `[UNK]`, and 99 unused tokens. Lastly, we split off a test set of 15,231 patients for fine-tuning applications and performed a 90:10 training-validation split on the remaining data.

B. Model

We used the Bidirectional Encoder Representations from Transformers (BERT) [13] transformer architecture, specifically the BERT-base specification with 768-dimensional embeddings, 12 attention heads, 12 transformer blocks. We used 0.1 dropout, parameter initialization of $\mathcal{N}(0, 0.02)$, and a post-LayerNorm transformer-block architecture. We used a sequence length of 128 and truncated extra tokens. Similar to the original BERT paper, we used word embeddings for the tokens and a learned sequence-position embedding [13]. We also added visit embeddings similar to [14], where all codes assigned during a particular hospital admission used the same visit embedding, based on visit order. Special tokens used a visit token of 0. The three embedding types were summed before being processed by the transformer model.

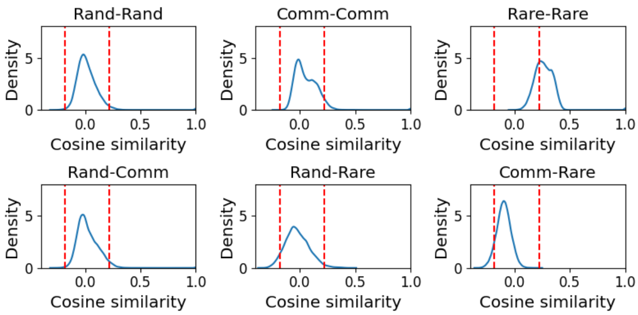

Fig. 1: Density of cosine similarity values by code frequency groups. Dashed lines represent the mean ± 2 std for the Rand-Rand distribution.

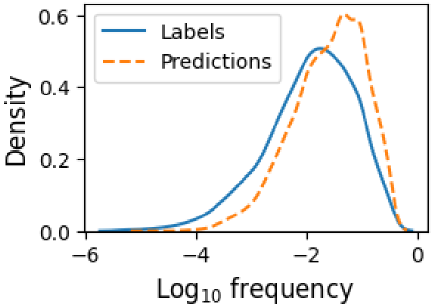

Fig. 2: Distributions of code frequencies in MLM labels (EHR data) and model predictions.

C. Training

Transformers are typically pre-trained on large unlabelled datasets using a self-supervised objective, before fine-tuning for a particular application on a smaller human-labelled dataset. Pre-training prepares the model for fine-tuning and largely determines the embeddings. BERT introduced the Masked Language Modeling (MLM) pre-training task, where the model is trained to predict hidden words in text. The architecture and MLM pre-training approach used for BERT has been adapted for medical applications in models such as BEHRT [5] and Med-BERT [14]. Instead of processing sequences of words, these models process sequences of diagnosis codes for fine-tuning tasks such as disease prediction.

We pre-trained our network with the MLM task using a masking probability of 15% and the same probabilities for assigning a `[MASK]` token, a random token, and the same token as in [13]. We pre-trained our network for 25 epochs on the training dataset with 157,401 patients. To investigate scaling, we also pre-trained our network for 25 epochs on subets of 25%, 50%, and 75% of the training dataset. We trained three separate network instances with different randomly initialized parameters. We use the Adam optimizer [15] with learning rate 1e-4, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate was scheduled with linear warm-up for the first 10% of steps and cosine decay for the remainder as in [13]. The training was performed on a single Nvidia RTX 3090 GPU. We used the Hugging Face transformers library for the PyTorch BERT model implementation and training [12].

III. Results

Based on the embeddings of the pre-trained BERT model, we generated t-SNE dimension reductions (not shown) to visualize how the ICD codes were clustered. Our projection showed three large lobes, with the most common codes all clustered at the tip of one lobe.

To supplement the insights from the t-SNE visualization, we carried out a small case study to understand which codes' embeddings had the highest cosine similarity with those of several common codes (Table 1). We observed some groupings of medically related codes with appropriately similar embeddings, such as a similarity score of 0.4470 between ICD-10 codes O99212 and O99213, representing obesity complicating pregnancy in the second and third trimester, respectively. However, some codes with similar embeddings had no obvious medical relation, neither associative nor syn-

Table 1: Case examples of ICD codes and cosine similarities

Code 1 (frequency in dataset)	Code 2	Case (frequency in dataset)	Cosine Similarity	Clinical relationship
J45909-10: Unspecified asthma, uncomplicated (4.212%)	E039-10: Hypothyroidism, unspecified	Common (5.05%)	0.313	No apparent link other than both having some immune system etiology.
	45991-10: Cough variant asthma	Rare (0.01%)	-0.155	These would be expected to be very similar.
78659-9: Other chest pain (3.612%)	7840-9: Headache	Common (2.372%)	0.315	Other than both being forms of pain, they appear unrelated.
	R071-10: Chest pain on breathing	Rare (0.006%)	-0.073	Sub-types of the same symptom; similarity would be expected
73300-9: Osteoporosis, unspecified (3.264%)	49390-9: Asthma, unspecified type, unspecified	Common (5.52%)	0.264	They seem unrelated.
	M818-10: Other osteoporosis without current pathological fracture	Rare (0.044%)	0.074	These would be expected to be identical.

onymous.

We observed that many high-similarity codes also had a relatively high frequency in the dataset (with the 300 most common codes appearing in 1.15 - 27.55% of the dataset). In contrast, we identified some strongly medically related codes that had a low cosine similarity. These low-similarity codes tended to also be rare in the dataset, appearing in fewer than 30 patient records in the pre-processed dataset while accounting for 72.17% of unique codes.

We then created groupings of diagnosis codes to compare by frequency, computed as the fraction of total patient records the code appeared in. We selected the 300 most common codes (Comm), a random sample of 300 rare codes that appeared in fewer than 30 patients (Rare), and a random sample of 300 codes from the whole dataset, weighted by frequency (Rand). The uncommon codes in Table 1 fall into the rare category. We computed cosine similarities between embedding vectors of codes within each group and between groups. The distributions of cosine similarities are shown in Figure 1. The means of each distribution were all significantly different from one another (two-tailed t-test corrected for multiple comparisons, $|z| > 10$, $p < 0.001$, $n = 90,000$). Table 2 presents the summary statistics of each distribution.

The cosine similarities between pairs of common codes (Comm-Comm) were significantly higher on average than those of pairs of random codes (Rand-Rand). Moreover, the Comm-Rare distribution had a significantly lower mean than the Rand-Rand distribution. The Rare-Rare distribution had a significantly higher mean than the Rand-Rand distribution as well.

The distributions in Figure 1 suggest suggest that the cosine similarity between medical code embeddings was influenced by the frequency of the code in the dataset, including spuriously high similarities among both common codes and rare codes, and low similarities between medically related common and rare codes. Without frequency-biased similarity between embedding vectors, we would expect the distribution of similarities in all these cases to be like Rand-Rand, where the bulk of codes have a cosine similarity close to 0, and a positive tail due to codes that are medically related to each other, which we would desire to have a high cosine similarity. We observed similar patterns in the embeddings of models trained on smaller subsets of the training data. Though we wondered whether the excessive similarity among rare codes might be more pronounced in smaller datasets, the mean Rare-Rare similarity was not higher after training with any of these subsets than with the full dataset, suggesting a weak relationship with dataset size.

Finally, we observed that the model's output was biased towards predicting more common codes. Figure 2 shows that the distribution of predicted codes had a higher density at higher frequency codes and did not reproduce the long tail of less frequent codes that was found in the data. The median log frequencies in the labels and predictions were -1.847 and -1.523 and the two distributions differed significantly (Mann–Whitney $U = 1,017,634,615$, $p < 0.001$, two-tailed,

Table 2: Cosine similarity distribution statistics

Distribution	Mean (std)	95% CI on the mean
Rand–Rand	0.018 (0.101)	[0.017, 0.018]
Comm–Comm	0.058 (0.109)	[0.057, 0.058]
Rare–Rare	0.264 (0.087)	[0.263, 0.264]
Rand–Comm	0.024 (0.098)	[0.024, 0.025]
Rand–Rare	-0.023 (0.108)	[-0.024, -0.023]
Comm-Rare	-0.098 (0.064)	[-0.099, -0.098]

$n = 51,983$).

IV. Discussion

The learned embedding similarities of diagnostic codes exhibited a systematic bias related to code frequency. Rare codes tended to be similar to each other, and the same was true for common codes. We also observed a bias in the network's output towards predicting high frequency codes. Past work on medical embeddings in language models suggested that embedding vectors for medically related concepts have high cosine similarity [5]. However, that study only examined the similarity neighbourhoods of the most common codes. Our results show that dissimilar vectors can be learned for medically similar codes with different frequencies.

Study limitations include that the dataset used was relatively small compared to those used by other BERT-based models such as BEHRT [5] (190 thousand vs. 1.6 million patients). Coding practices may also differ by hospital; codes that are used infrequently in one setting may not actually describe an uncommon event. Future work should evaluate whether the biases found occur with other medical datasets.

In conclusion, this study suggests that though this frequency bias may improve accuracy on MLM tasks, BERT transformer models trained on health record data may generalize poorly to less frequently used medical codes. This is problematic because the distribution of code frequencies has a long tail, with rare codes more likely to be confused for each other and less likely to be predicted overall (Figure 2). This motivates further study towards improving generalization.

Conflict of Interest

The authors declare that this research was sponsored by Smile Digital Health.

References

1. E. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019.
2. A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, May 2018. [Online]. Available: https://doi.org/10.1038/s41746-018-0029-1
3. E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl, and N. H. Shah, "Language models are an effective representation learning technique for electronic health record data," *Journal of Biomedical Informatics*, vol. 113, p. 103637, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420302653
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762
5. Y. Li, S. Rao, J. R. A. Solares, A. Hassaïne, D. Canoy, Y. Zhu, K. Rahimi, and G. S. Khorshidi, "BEHRT: transformer for electronic health records," *CoRR*, vol. abs/1907.09538, 2019. [Online]. Available: http://arxiv.org/abs/1907.09538
6. B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," *CoRR*, vol. abs/2011.05864, 2020. [Online]. Available: https://arxiv.org/abs/2011.05864
7. C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T. Liu, "FRAGE: frequency-agnostic word representation," *CoRR*, vol. abs/1809.06858, 2018. [Online]. Available: http://arxiv.org/abs/1809.06858
8. Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, and K. Rahimi, "Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," 2021. [Online]. Available: https://arxiv.org/abs/2106.11360
9. A. Ashfaq, M. Lingman, and S. Nowaczyk, "Kafe: Knowledge and frequency adapted embeddings," in *Machine Learning, Optimization, and Data Science: 7th International Conference, LOD 2021, Grasmere, UK, October 4–8, 2021, Revised Selected Papers, Part II*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 132–146. [Online]. Available: https://doi.org/10.1007/978-3-030-95470-3_10
10. A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv (version 2.0," 2022. [Online]. Available: https://doi.org/10.13026/7vcr-e114
11. A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.
12. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2019. [Online]. Available: https://arxiv.org/abs/1910.03771
13. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805
14. L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction," *CoRR*, vol. abs/2005.12833, 2020. [Online]. Available: https://arxiv.org/abs/2005.12833
15. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980