Med-BERT: pretrained contextualized embeddings on largescale structured electronic health records for disease prediction(npj digital medicine 2021)

Author : Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao & Degui Zhi

Paper Link : https://www.nature.com/articles/s41746-021-00455-y

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/b79d1aa1-4fa9-4f1f-b93b-6d0e670921af/MED_BERT(2021).pdf

Code: https://github.com/ZhiGroup/Med-BERT

1. Abstract

EHR(전자 건강 기록)을 이용한 딥러닝 기반 예측 모델은 많은 임상 업무에서 매우 인상적인 성
능을 보입니다. 하지만 이러한 모델은 높은 정확도를 위해 대규모 훈련 집단이 필요한 경우가 많
아 딥러닝 기반 모델의 채택에 어려움이 있습니다. 최근 들어, 자연어 처리 분야에서 BERT는
매우 큰 성공을 거뒀습니다. 대규모 훈련 corpus에 대한 BERT 사전 학습은 더 작은 데이터에
서 훈련된 모델의 성능을 향상시킬 수 있는 맥락별 임베딩을 생성합니다.
저자는 BERT로부터 영감을 받아 BERT 프레임워크를 구조화된 EHR 도메인에

‘MedBERT’를
제안합니다. Med-BERT는 28,490,650명 환자의 EHR 데이터로부터 사전 학습된 상
황별 임베딩 모델입니다.
Fine-tuning을 통해 Med-BERT는 예측 정확도를 크게 향상시켰고, 두 가지 질병 예측 작업
에
서 AUC 곡선 아래 영역을 1.21~6.14% 확장했습니다. 사전 학습된 Med-BERT는 특히
FineTuning된
훈련 데이터에 대해 좋은 성능을 얻으며, Med-BERT를 적용하지 않은 딥러닝 모델과
비교해 20% 이상의 AUC를 얻고, 10배 더 큰 훈련 데이터로 훈련한 모델 만큼 높은 AUC를
얻
을 수 있었습니다.
Med-BERT가 소규모 로컬 훈련 데이터를 사용하면 질병 예측 연구에 기여하고, 데이터 수
집 비
용을 절감하고, AI 지원 의료 속도를 가속화할 것으로 기대합니다.

2. contribution

1. EHR에 대한 이전 연구는 word2vec, Glove 이였지만, 이는 Context를 고려하지 않고 Embedding을 수행하기 때문에, Context의 순서까지 고려할 수 있는 BERT채택
2. 전자 의료 기록은 text와 유사하기에, transformer의 bidirectional encoder 구조로 nlp 분야의 많은 발전을 가져온 pre-training BERT에서 EHR을 이용해 fine tuning 시키면 EHR기반 predictive modeling의 성능을 향상시킬 수 있음
3. MedBERT는 기존의 BEHRT, G-BERT에 비하여 훨씬 큰 vocabulary size와 pretraining cohort size를 가짐

Table 1. Comparison of Med-BERT with BEHRT and G-BERT from multiple perspectives.

Criteria	BEHRT	G-BERT	Med-BERT
Type of input code	Caliber code for diagnosis developed by a college in London	Selected ICD-9 code for diagnosis + ATC code for medication	ICD-9 + ICD-10 code for diagnosis
Vocabulary size	301	<4K	82K
Pre-training data source	CPRD (primary care data) [45]	MIMIC III (ICU data) [46]	Cerner HealthFacts (general EHRs)
Input structure	Code + visit + age embeddings	Code embeddings from ontology + visit embeddings	Code + visit + code serialization embeddings
Pre-training sample unit	Patient's visit sequence	Single visit	Patient's visit sequence
Total number of pre-training patients	1.6M	20K	20M
Average number of visits for each patient for pre-training	Not reported but > 5	<2	8
Pre-training task	Masked LM	Modified Masked LM	Masked LM + prediction of prolonged length of stay in hospital
Evaluation task	Diagnosis code prediction in different time windows	Medication code prediction	Disease predictions according to strict inclusion/exclusion criteria
Total number of patients in evaluation tasks	699K, 391K, and 342K for different time windows	7K	50K, 20K, and 20K for three task cohorts

4. 기존의 BEHRT, G-BERT와 에서는 방문 순서를 고려하지 않았지만, **Med-BERT에서는 serialization embedding을 추가하여 병원 방문 순서까지 고려할 수 있음.**
5. pretrained Med-BERT는 당뇨병(DHF), 심부전(PaCa)을 예측하는 2개의 task에 의해 fine-tunning함(Cerner Health Facts, Truven Health MarketScan **EHR 데이터 베이스**)

3. Method

1. 학습 데이터 준비

- Cerner Health Facts, Truven Health MarketScan (약 2800만 개의 EHR row data)

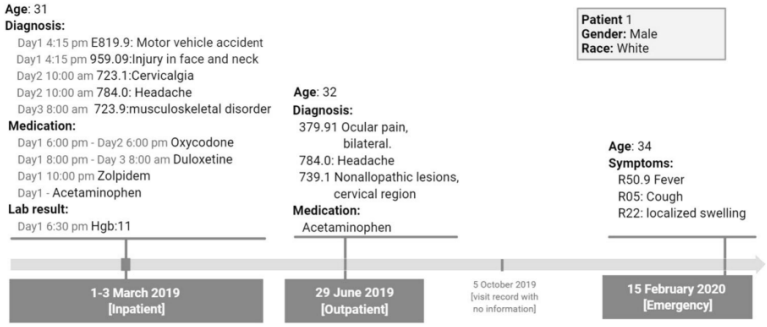
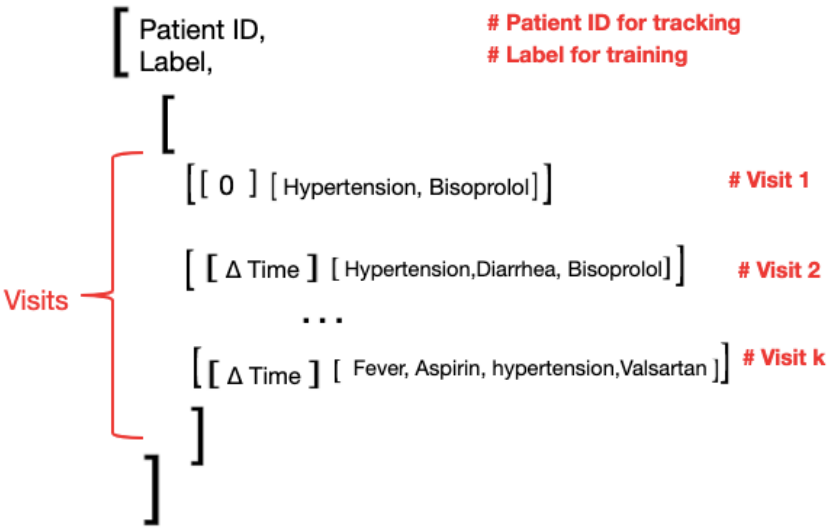


Fig. 2 An example of structured EHR data of a hypothetical patient as it would be available from a typical EHR system (e.g., Cerner or Truven). For this patient, four visits with dates and encounter types are organized according to chronological order at the bottom. Detailed information including demographic and medical codes with time stamps are shown above. Note that not all information is recorded, as in real-world EHR recording system.

- data structure



- example data

patient_id	admission_date	discharge_date	Medical_code	present_on_admit	diagnosis_priority	Billing_Source
pt_1	2014-03-01	2014-03-01	ICD9_428.1	0	1	0
pt_1	2014-03-01	2014-03-01	ICD9_723.1	0	2	0
pt_1	2015-03-07	2015-03-10	ICD10_I50.41	0	1	0
pt_2	2016-07-18	2016-07-21	ICD9_453.4	1	1	0
pt_2	2016-07-18	2016-07-21	ICD9_415.1	0	1	1

2. BERT vs Med-BERT 의 Architecture 비교

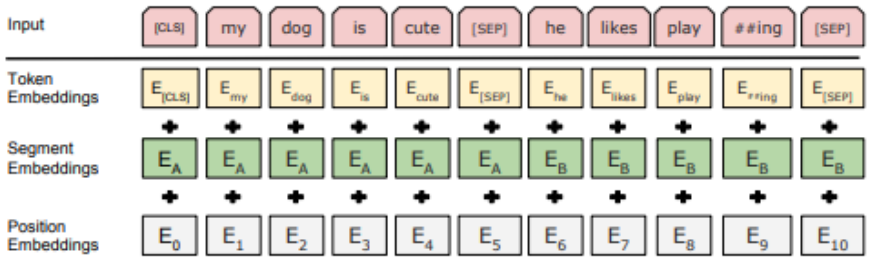


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

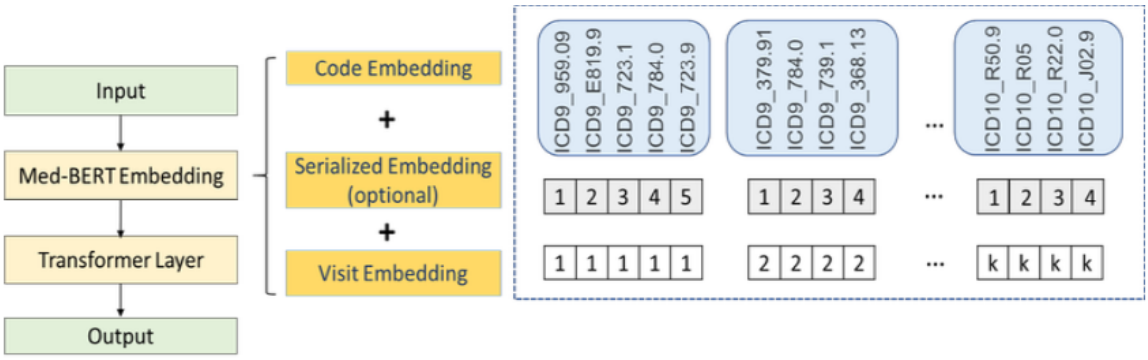


Figure 2. Med-BERT structure.

- 의학 단어를 masking, 검사수치, 숫자 위주로 masking,
- clinical bert dataset, mimic dataset → 요청을 받아야 한다
- visit embedding : 응급환자가 온 시간, 오전 오후 저녁에 대한 정보를 embedding으로 넣는 것 도 생각
- indicator가 합쳐진 형태
- input Token
  - Input layer에서 Special token을 사용하지 않는다.
    1. BERT와의 Input_dim을 맞추기 위함
    2. **[CLS] : EHR sequance 일반적으로 일반 문장에 비해 매우 길기 때문에 , [CLS] token을 이용하면 정보 손실이 크다 ex. 환자가 10번 방문한다면 매우 Sentence는 매우 길어진다**

       → Feed Fowrad Network를 추가
- Embedding
  - code Embedding : Token으로부터 각 임상 코드를 나타낸다(ICD-10,ICD9)

    - **serialization embedding : 방문의 순서를 고려할 수 있게 해줌 [1,2,3,4,5] 번째 방문시의 ICD code로 날짜별, 진단내용을 표현할 수 있다**
    - **Visit Embedding : 방문 했을때를 구별해 줌**
- Parameter setting
    - (L=6, A=6 , H=192) cf. L : Encoder Block / A: Attention head / H: hidden layer
    - maximum sequence length : 521 tokens
    - optimizer : Adam & learning rate : 5e-5 & drop out rate : 0.1
- Masked Language modeling
    - 80[masked]-10[random token]-10[original token]
- Transformer layer

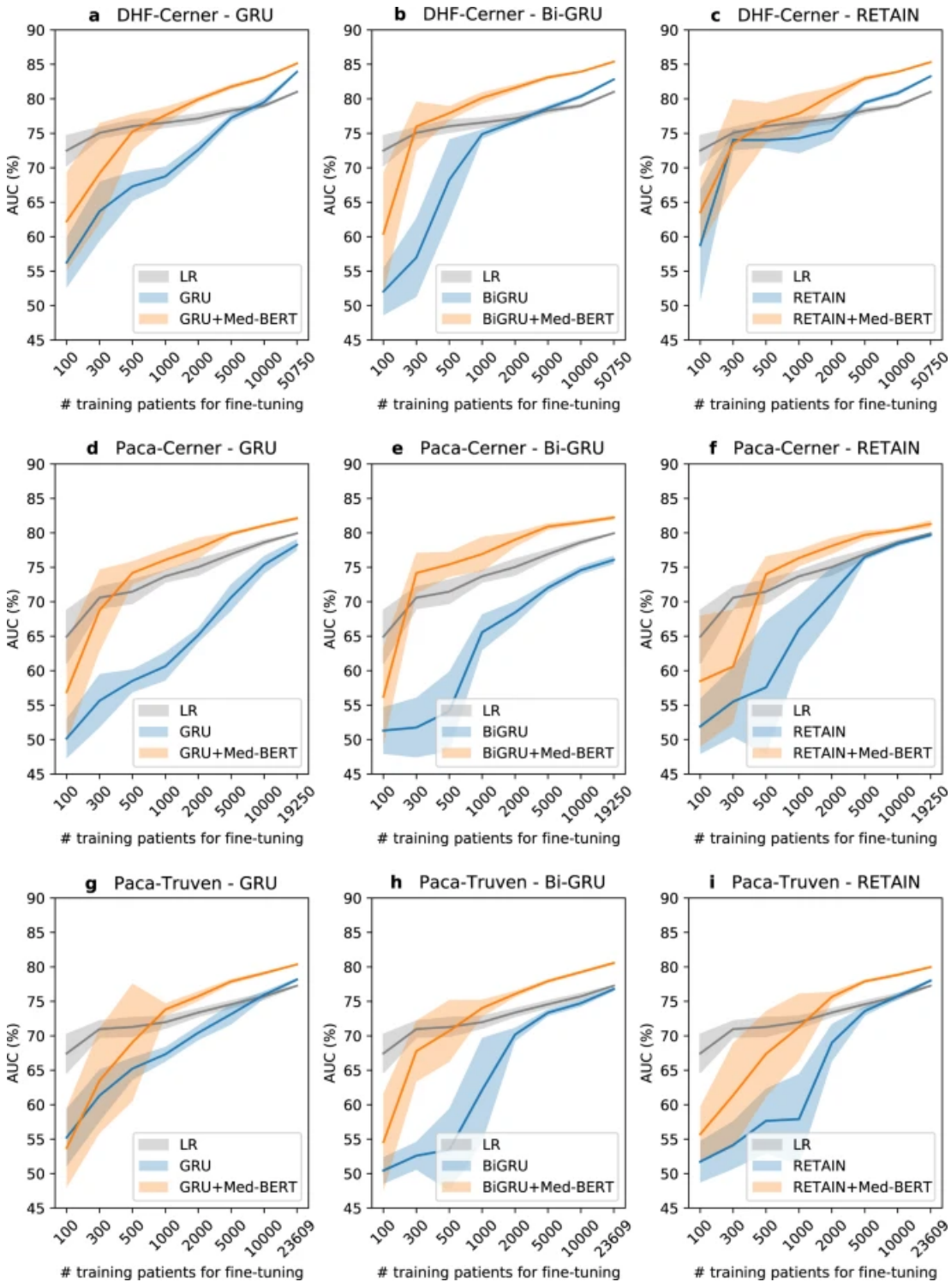
  기존의 Transformer layer와 동일

4. Experiments

pretrained Med-BERT vs untrained Med-BERT

- experiment setup
    - pretrained Med-BERT : figure2
    - untrained Med-BERT : randomly initialized token + segment emb + positional emb
- result

  아래는 성능을 나타내는 AUC 테이블과 플롯입니다.



- 훈련되지 않은 Med-BERT가 Med-BERT보다 성능이 훨씬 좋지 않음 → 사전 학습 단계가 향상된 성능에 더 중요한 역할
- 훈련되지 않은 Med-BERT는 매우 큰 수의 매개 변수화된 모델이므로 과적합의 위험 있음
- 반면, 사전 학습된 모델은 매우 큰 데이터에 robust한 구성으로 시작되었기 때문에 일반화

- DHF-Cerner -GRU를 보면, Med-BERT가 없이 GRU만 사용할 경우 샘플이 1,000개 미만일 때 AUC가 겨우 0.65를 넘는 수준임을 확인
- Med-BERT를 적용하면 AUC가 20% 이상 증가
- PaCa-Cerner의 경우 거의 모든 훈련 크기에 대해 GRU와 Bi-GRU에 Med-BERT를 추가함으로써 큰 개선이 입증
- 샘플의 크기가 1,000을 초과할 때 모든 예측에서 Med-BERT가 가장 좋은 성능

5. 사견

- 사전 학습에 사용한 데이터는 구조화 된 EHR 데이터로 모두 딕셔너리처럼 “A:숫자”와 같은 형식으로 저장되고 불러오기 때문에 현재 보유중인 데이터셋의 전처리 방향 결정 필요
- 본 논문의 task는 환자인지 아닌지를 분류하는 task인데, 병인 경우와 아닌 경우의 threshold 조절을 통해 응급 여부로 이을 수 있을 것으로 기대
- 본 논문에 나온 embedding 방식처럼 serialization embedding과 visit embedding을 추가한다면, 시간 순서에 따른 환자 상태 정보도 사용할 수 있을 것으로 기대
- 본 논문에서는 [SEP] token을 pretrain 과정에서 classification 작업을 위해

  Q&A방법을 사용하는 대신, pretrain dataset에 대해 상대적으로 발병률이 높고(대중적인 병)을 가진 feature를 선택했다.
  이에, 가장 일반적으로 사용되는 의료 지표인 사망률, 조기재입원 맟 장기간 방문(LOS>7일) 인 경우를 가진 feture를 선택하는게 pretrain 단계에서 더 좋은 성능을 보여주었다는데,

  우리의 task에서도 이와 비슷하게 위험한 수치(ex. WBC,온도)가 높게 측정된 정보를 이용해 응급한지를 예측할 수 있을 것 으로 기대

  또는, 우리의 task에서는 [SEP] token을 응급한지 아닌지 결정할 수 있게 downstream task로 바꿀 수 있을 것으로 기대