Enriching Unsupervised User Embedding via Medical Concepts

Xiaolei Huang[1], Franck Dernoncourt[2], Mark Dredze[3]

1. University of Memphis 2. Adobe Research 3. Johns Hopkins University

1. Unsupervised User Embedding is effective to capture patient patterns with broad real-world applications.

2. Medical concepts significantly boost patient modeling & potentially enhance interpretation.

https://github.com/xiaoleihuang/UserEmb_Explainable

User embedding models user behaviors by mapping all user info into a unified vector space.

Medical concepts: basic units for medical info, such as disease symptom and clinical drug.

Unsupervised

avoid *error-prone labels*[1]

reduce *labor*

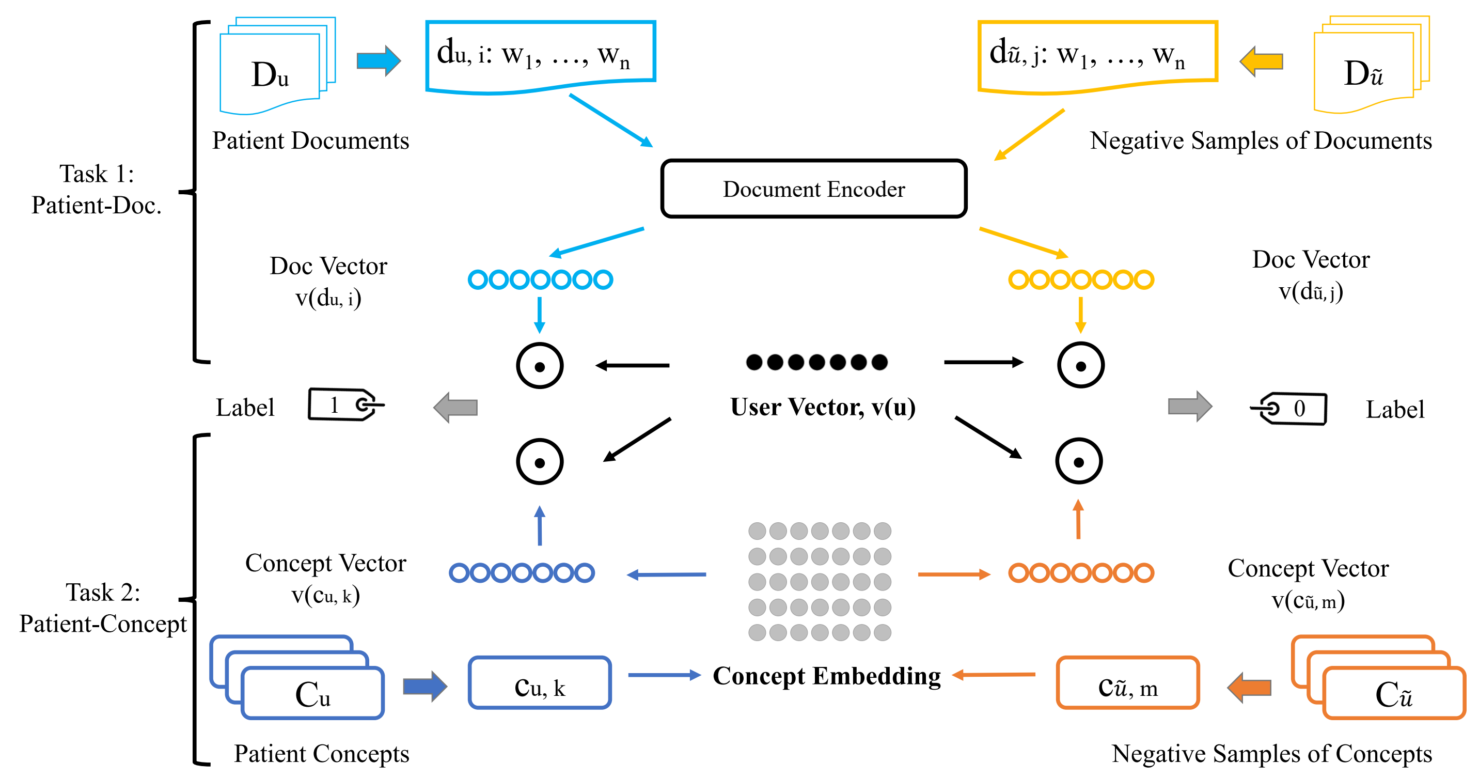
save *money*

Data	Statistics			
	Doc	User	U-label	U-Concept
Diabetes[3]	1265	288	10	89
MIMIC-III[2]	54888	48807	276	94

MetaMap[4]

Model

Concept-Aware User Embedding (*CAUE*)



Task 1: Patient-Document	Task 2: Patient-Concept
$\mathcal{L}(u,d) = -log(\sigma(v(u)\cdot v(d_u))) - log(1-\sigma(v(u)\cdot v(d_{\tilde{u}})))$	$\mathcal{L}(u,c) = -log(\sigma(v(u)\cdot v(c_u))) - log(1-\sigma(v(u)\cdot v(c_{\tilde{u}})))$

simulate diagnosis process:
enforce models to recognize patients (u) of medical notes (d) / concepts (c).

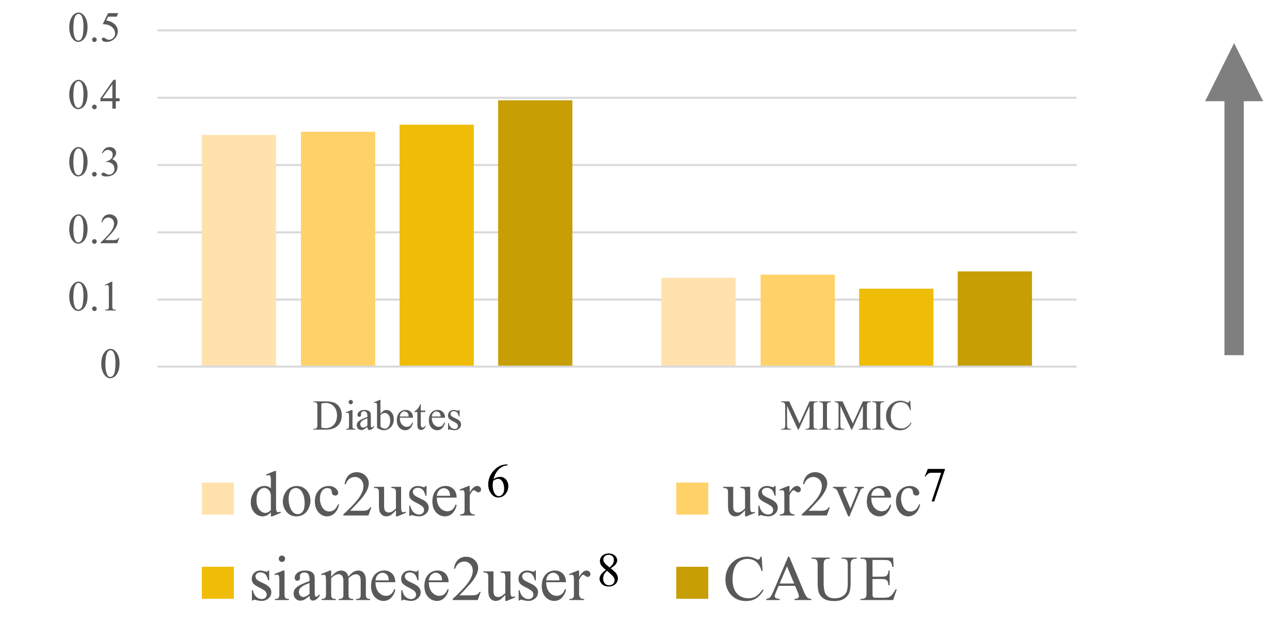
Key Methods

1. *contrastive learning*[5]: generate counterfactuals → model robustness
2. *negative sampling*: convert to binary prediction (self-supervision)

Eval & Apps

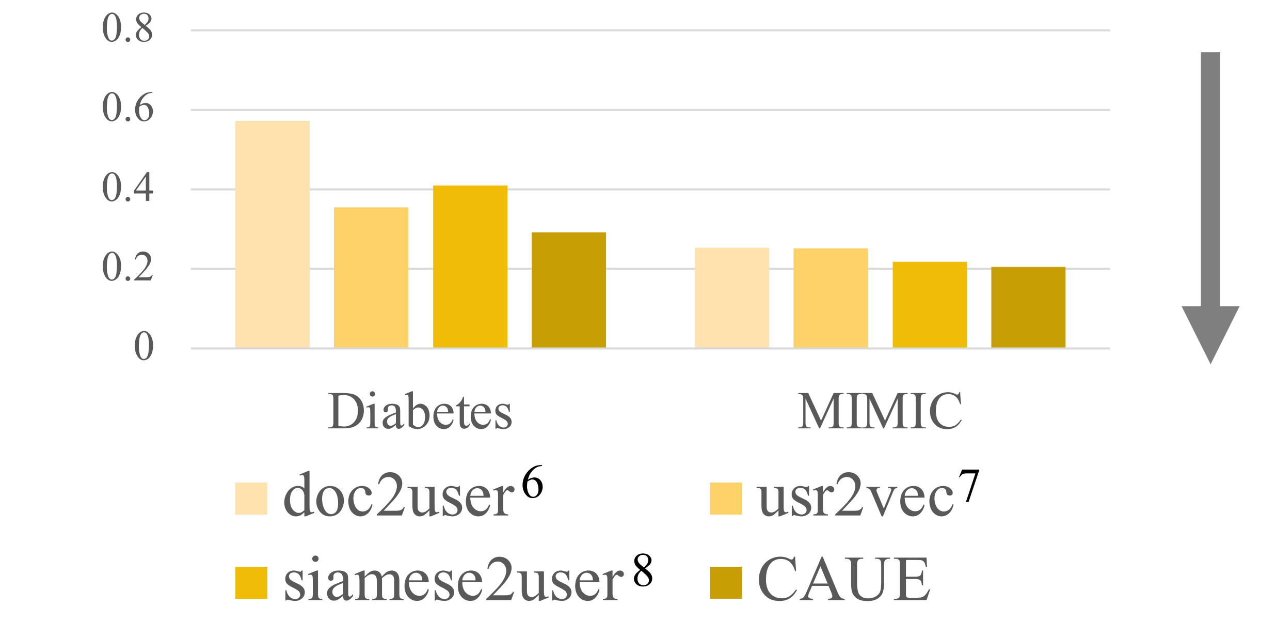
Patient Retrieval

$$Jaccard(u_1,u_2) = \frac{|l_1 \cap l_2|}{|l_1 \cup l_2|}$$



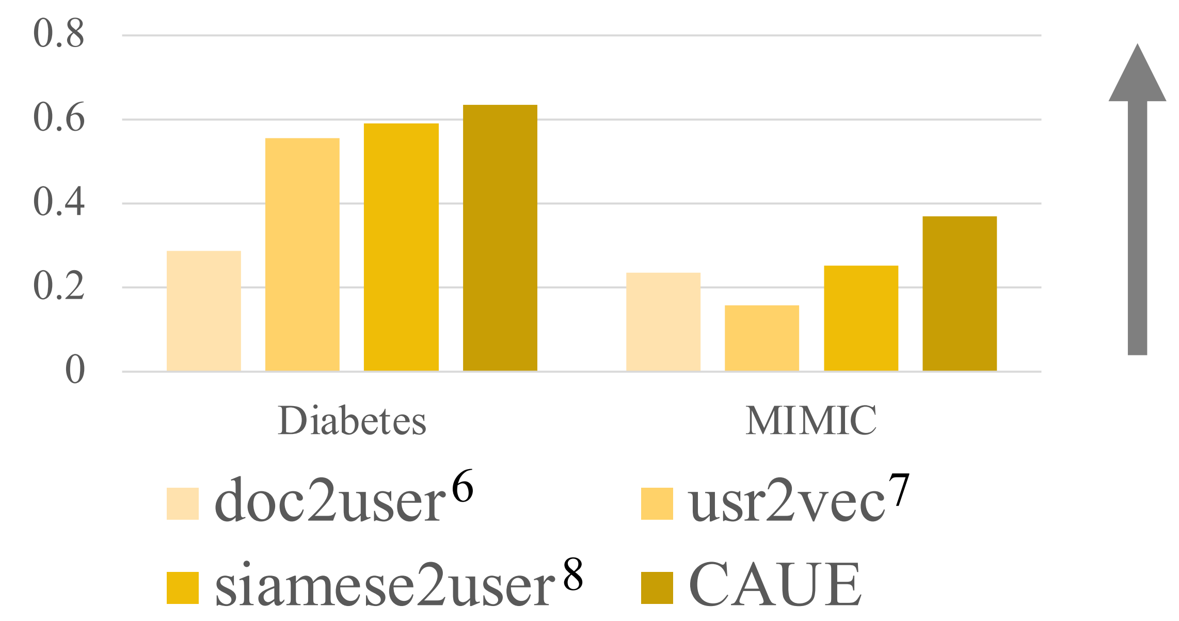
Patient Relatedness

MSE: Mean Square Error



Phenotype Inference

MAP: Mean Average Precision



Reference:

1. Birman-Deych, et al. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors.
2. Johnson, et al. Mimic-iii, a freely accessible critical care database.
3. Stubbs et al. Cohort selection for clinical trials: n2c2 2018 shared task track 1.
4. Aronson and Lang. An overview of MetaMap: historical perspective and recent advances.
5. Logeswaran and Lee. An efficient framework for learning sentence representations.
6. Ding, et al. Predicting delay discounting from social media likes with unsupervised feature learning.
7. Amir et al. Quantifying mental health from social media with neural user embeddings
8. Mueller and Thyagarajan. Siamese recurrent architectures for learning sentence similarity.