GRAM: **Gr**aph-based **A**ttention **M**odel for Healthcare Representation Learning (2017)

by

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, Jimeng Sun

Presentation by:
Miguel Cardoso 2020

Deep Learning in Healthcare

Deep Learning in Healthcare

- Exhibit promising performance for predictive modeling in various domains
- Can only predict what's on the training set
- Need balanced dataset
- Hard to understand the reasoning behind the predictions

Deep Learning in <mark>Healthcare</mark>

- A lot of volume of data but usually sparse
- Need to be able model diseases not in the dataset
- Needs interpretability

Deep Learning in Healthcare

- Word Embeddings
- CNNs
- SDA
- RNNs

EHR   Prediction

All good and games but data is still an issue

Add Domain-Knowledge

- Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT)
- International Classification of Diseases (ICD)
- Clinical Classification Software (CCS)
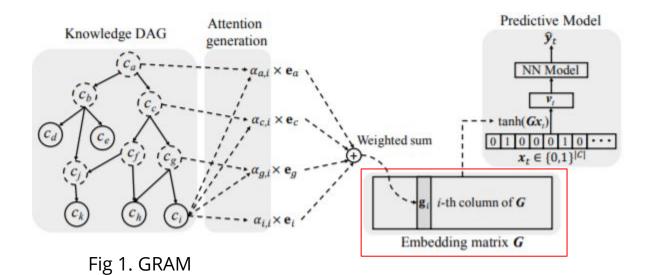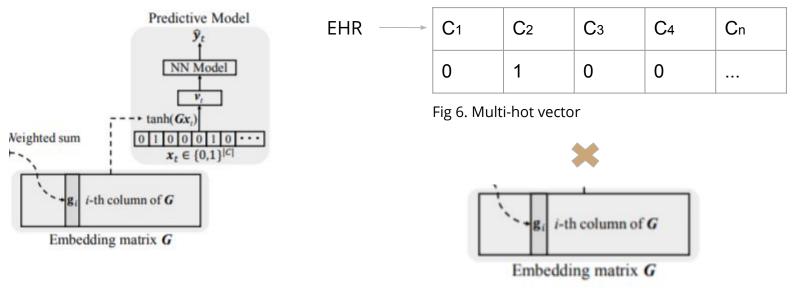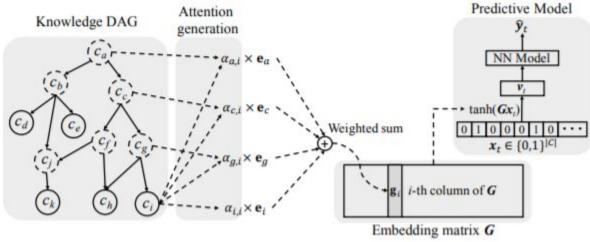
Ontologies are our best friend

GRAM - **Gr**aph-based **A**ttention **M**odel

Combine Domain-Knowledge with Deep Learning models via attention mechanisms

Fig 1. GRAM

GRAM - **Gr**aph-based **A**ttention **M**odel

Combine Domain-Knowledge with Deep Learning models via attention mechanisms

Fig 1. GRAM

GRAM - **Gr**aph-based **A**ttention **M**odel

Fig 2. Predictive Model

EHR

C₁	C₂	C₃	C₄	Cₙ
0	1	0	0	...

Fig 6. Multi-hot vector

Given v₁...vₜ let's predict vₜ₊₁ disease codes

Combine Domain-Knowledge with Deep Learning models via attention mechanisms
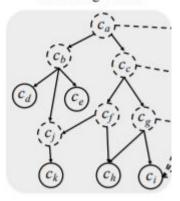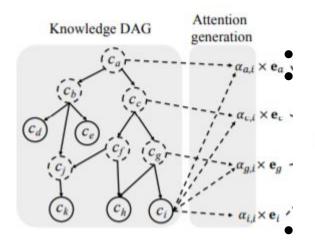

Fig 3. Knowledge DAG

- Parent-Child relationship
- Leaf nodes are concepts in the EHR
- When a medical concept is less frequent in the data, more weight is given to its ancestors

Combine Domain-Knowledge with Deep Learning models via attention mechanisms

Fig 4. Knowledge DAG and Attention Generation

- Each node is assigned a basic embedding vector $e_i \in R^m$
- A leaf node's final representation is a convex combination of the basic embeddings of itself and its ancestors.

$$g_i = \sum_{j \in \mathcal{A}(i)} \alpha_{ij} e_j, \qquad \sum_{j \in \mathcal{A}(i)} \alpha_{ij} = 1, \ \alpha_{ij} \geq 0 \ \text{for } j \in \mathcal{A}(i), \quad (1)$$

$$\alpha_{ij} \in \mathbb{R}^+$$

- GloVe[1] is used to setup the initial basic embeddings

[1] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In EMNLP

Combine Domain-Knowledge with Deep Learning models via attention mechanisms

- **g**i is the weighted sum of Ci and its ancestors
- **G** is the concatenation of all representations g1 ,.. g|C|

Fig 5. Representation Learning

GRAM - **Gr**aph-based **A**ttention **M**odel

Combine Domain-Knowledge with Deep Learning models via attention mechanisms

Fig 1. GRAM

GRAM - **Gr**aph-based **A**ttention **M**odel

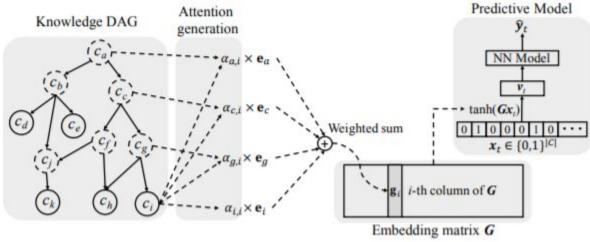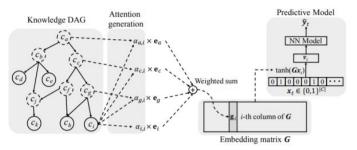
Fig 1. GRAM

Experiments

- Prediction performance when facing data insufficiency
  - Sequential Diagnosis
  - Heart Failure prediction

Model Assessment

Accuracy@k for each of CCS single-level diagnosis codes such that, given a visit Vt , we get 1 if the target diagnosis is in the top k guesses and 0 otherwise.

Binary prediction task for predicting a future HF onset where the prediction is made only once at the last visit.

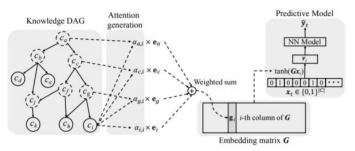Try different levels of data insufficiency.

GRAM - **Gr**aph-based **A**ttention **M**odel

Fig 1. GRAM

Datasets

- Sutter Palo Alto Medical Foundation (PAMF), 10-years longitudinal medical records of 258K primary care patients between 50 to 89 years of age
- MIMIC-III (7.5K intensive care unit (ICU) patients over 11 years)
- Sutter heart failure (HF) cohort, which is a subset of Sutter PAMF data for a heart failure onset prediction study with 3.4K HF cases

Table 1: Basic statistics of Sutter PAMF, MIMIC-III and Sutter heart failure (HF) cohort.

Dataset	Sutter PAMF	MIMIC-III	Sutter HF cohort
# of patients	258,555†	7,499†	30,727† (3,408 cases)
# of visits	13,920,759	19,911	572,551
Avg. # of visits per patient	53.8	2.66	38.38
# of unique ICD9 codes	10,437	4,893	5,689
Avg. # of codes per visit	1.98	13.1	2.06
Max # of codes per visit	54	39	29

† For all datasets, we chose patients who made at least two visits.

GRAM - **Gr**aph-based **A**ttention **M**odel

Experiments

- Prediction performance when facing data insufficiency
  - Sequential Diagnosis
  - Heart Failure prediction

Model Assessment

Accuracy@k for each of CCS single-level diagnosis codes such that, given a visit V_t , we get 1 if the target diagnosis is in the top k guesses and 0 otherwise.

Binary prediction task for predicting a future HF onset where the prediction is made only once at the last visit.

Model	0-20	20-40	40-60	60-80	80-100
GRAM+	**0.0150**	**0.3242**	0.4325	0.4238	0.4903
GRAM	0.0042	0.2987	0.4224	0.4193	0.4895
RandomDAG	0.0050	0.2700	0.4010	0.4059	0.4853
RNN+	0.0069	0.2742	0.4140	0.4212	**0.4959**
RNN	0.0080	0.2691	0.4134	0.4227	0.4951
SimpleRollUp	0.0085	0.3078	**0.4369**	**0.4330**	0.4924
RollUpRare	0.0062	0.2768	0.4176	0.4226	0.4956

(a) *Accuracy@5 of sequential diagnoses prediction on Sutter data*

Model	0-20	20-40	40-60	60-80	80-100
GRAM+	**0.0672**	**0.1787**	**0.2644**	0.2490	0.6267
GRAM	0.0556	0.1016	0.1935	0.2296	0.6363
RandomDAG	0.0329	0.0708	0.1346	0.1512	0.4494
RNN+	0.0454	0.0843	0.2080	0.2494	0.6239
RNN	0.0454	0.0731	0.1804	0.2371	0.6243
SimpleRollUp	0.0578	0.1328	0.2455	**0.2667**	**0.6387**
RollUpRare	0.0454	0.0653	0.1843	0.2364	0.6277

(b) *Accuracy@20 of sequential diagnoses prediction on MIMIC-III*

Model	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
GRAM+	0.7970	**0.8223**	0.8307	**0.8332**	**0.8389**	**0.8404**	**0.8452**	**0.8456**	**0.8447**	**0.8448**
GRAM	**0.7981**	0.8217	**0.8340**	**0.8332**	0.8372	0.8377	0.8440	0.8431	0.8430	0.8447
RandomDAG	0.7644	0.7882	0.7986	0.8070	0.8143	0.8185	0.8274	0.8312	0.8254	0.8226
RNN+	0.7930	0.8117	0.8162	0.8215	0.8261	0.8333	0.8343	0.8353	0.8345	0.8335
RNN	0.7811	0.7942	0.8066	0.8111	0.8156	0.8207	0.8258	0.8278	0.8297	0.8314
SimpleRollUp	0.7799	0.8022	0.8108	0.8133	0.8177	0.8207	0.8223	0.8272	0.8269	0.8258
RollUpRare	0.7830	0.8067	0.8064	0.8119	0.8211	0.8202	0.8262	0.8296	0.8307	0.8291
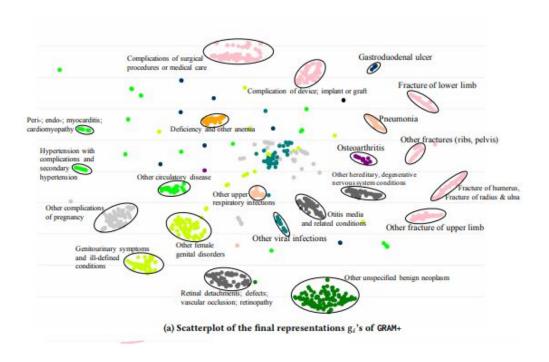
(c) *AUC of HF onset prediction on Sutter HF cohort*

GRAM - **Gr**aph-based **A**ttention **M**odel

Scalability

- 50% more time to train for a single epoch for all prediction tasks
- Attention Weights
- GRAM overhead
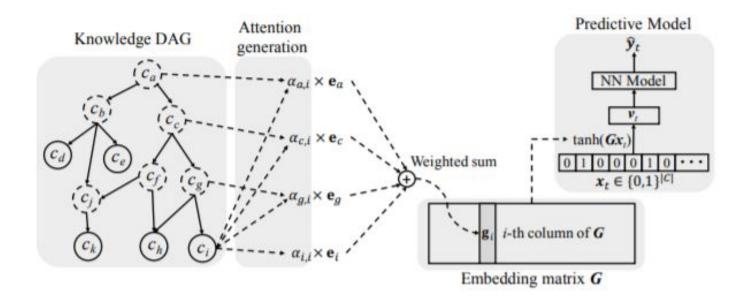- 50% more epochs to reach the model with the lowest validation loss

Table 3: Scalability result in per epoch training time in second (the number of epochs needed). SDP stands for Sequential Diagnoses Prediction

Model	SDP (Sutter data)	SDP (MIMIC-III)	HF prediction (Sutter HF cohort)
GRAM	525s (39 epochs)	2s (11 epochs)	12s (7 epochs)
RNN	352s (24 epochs)	1s (6 epochs)	8s (5 epochs)

GRAM - **Gr**aph-based **A**ttention **M**odel

(a) Scatterplot of the final representations g_i's of GRAM+

Related Work and References

- Doctor AI: Predicting Clinical Events via Recurrent Neural Networks
- RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism
- GRAM: Graph-based Attention Model for Representation Learning
- Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer
- Graph Neural Network-Based Diagnosis Prediction

Questions ?