KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare

Fenglong Ma¹, Quanzeng You², Houping Xiao³, Radha Chitta⁴, Jing Zhou⁵, Jing Gao¹

¹ University at Buffalo, ² Microsoft AI & Research  
³ Georgia State University, ⁴ Kira Systems, ⁵ eHealth Inc.

CIKM 2018
Background

- **Electronic Health Records**

  "An electronic health record (EHR), or electronic medical record (EMR), is the systematized collection of patient and population electronically-stored health information in a digital format."

  --- Wikipedia

- **Personalized Medicine**
Background

- **Electronic Health Records (EHR)**
  - A comprehensive EHR dataset that contains everything happened to a patient at the hospital.

- **Structured Codes**
- **Lab Measures**
- **Spectrograms**
- **Images**
- **Free Text**
Challenges of Mining EHR Data

- **EHR Data with Structured Codes**
  - Temporal
  - High dimensional
  - Noisy

An example of a patient’s visit information.
Task

- **Diagnosis Prediction**
  - Utilizing historical EHR data of individuals to predict the next visit information.
Existing Work

- Deep Learning based Diagnosis Prediction
  - Med2Vec

Med2Vec does not consider the temporal nature of EHR data.

Existing Work

- **Deep Learning based Diagnosis Prediction**
  - **RETAIN**

The predictive performance of RETAIN drops when the length of patient visits is large.

Existing Work

- Deep Learning based Diagnosis Prediction
  - Dipole

(1) Training RNN-based models requires large amounts of data.
(2) Medical codes of rare diseases may not learn correct embeddings.

- Ma et al. *Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks*. In KDD’17.
Existing Work

- **Deep Learning based Diagnosis Prediction**
- **GRAM**

Medical **ontology** information is only used when learning code representations, which does **not explicitly** affect the **final predictions**.

- Choi et al. **GRAM: Graph-based Attention Model for Healthcare Representation Learning**. In KDD’17.
Existing Work

- **Deep Learning based Diagnosis Prediction**
  - **GRAM**

When sufficient training data is available, GRAM has relatively comparable performance with other RNN variants such as Dipole.

Challenges

- How to train a model to deal with the problem of rare diagnosis codes?
- How to fully utilize medical knowledge graphs for diagnosis prediction task?
- How to design a robust predictive model?
- How to interpret the learned embeddings by the proposed model?
KAME: Knowledge-based Attention Model

- **Overview**

- **Novelties**
  - Take high-level visit information as input.
  - Propose a knowledge attention mechanism.
  - Consider general knowledge when making prediction.
KAME: Knowledge-based Attention Model

- **Inputs**
  - Knowledge graph $G = \{C, N\}$.
  - Visit record and Ancestor set (high-level)
  - $V_t = \{c_1, c_2, c_4\}$, $x_t = [1,1,0,1,0]$.
  - $Q_t = \{n_1, n_2, n_3, n_5\}$, $f_t = [1,1,1,0,1]$. 

Diagram showing the Knowledge Graph and Knowledge Attention model with nodes and edges representing the structure and flow of information.
KAME: Knowledge-based Attention Model

Knowledge Graph Embedding

- Graph Attention
- Leaf code $m_1 = \text{GRAM}(e_1, a_1, a_2)$
- Leaf Code Matrix $M$
- Ancestor Code Matrix $A$

Visit Embedding & Recurrent Neural Network

- Visit Embedding
  \[ v_t = \tanh(Mx_t) \]

- Gated Recurrent Unit (GRU)
  \[ h_t = \text{GRU}(v_t; \Omega) \]
- **Knowledge-based Attention Mechanism**
  
  - **Latent Knowledge Embedding**
    \[ \mathbf{L}_n^t = \theta(\mathbf{A}, \mathbf{f}_{tn}) = (\mathbf{W}_k \mathbf{A}_n + \mathbf{b}_k)\mathbf{f}_{tn} \]
  
  - **Example**: Ancestor set \( Q_t = \{n_1, n_2, n_3, n_5\} \), \( \mathbf{f}_t = \{1,1,1,0,1\} \), \( \mathbf{f}_{t1} = 1 \) and \( \mathbf{f}_{t4} = 0 \).
  
  - **Attention**
    \[ k_t = \sum_{n=1}^{\vert N \vert} \alpha_{tn} \mathbf{L}_n^t, \quad \alpha_{tn} = \frac{\exp(h_t^T \mathbf{L}_n^t)}{\sum_{j=1}^{\vert N \vert} \exp(h_t^T \mathbf{L}_n^t)} \]
### Knowledge-based Prediction

- **Knowledge Attentional Vector**
  \[ s_t = [h_t; k_t] \]

- **Prediction**
  \[ \hat{y}_t = \text{Softmax}(W_c s_t + b_c) \]
**REMARK**

- KAME is the **generalization** of the state-of-the-art diagnosis prediction model GRAM.
- When removing the proposed knowledge-based attention component (i.e., deleting $k_t$), then the proposed KAME is reduced to GRAM.
Experiments

- **Datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Medicaid</th>
<th>Diabetes</th>
<th>MIMIC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td># of patients</td>
<td>99,159</td>
<td>17,584</td>
<td>7,499</td>
</tr>
<tr>
<td># of visits</td>
<td>2,034,485</td>
<td>466,024</td>
<td>19,911</td>
</tr>
<tr>
<td>Avg. # of visits per patient</td>
<td>20.52</td>
<td>26.50</td>
<td>2.66</td>
</tr>
<tr>
<td># of unique ICD9 codes</td>
<td>9,701</td>
<td>7,437</td>
<td>4,880</td>
</tr>
<tr>
<td>Avg. # of ICD9 codes per visit</td>
<td>2.78</td>
<td>3.39</td>
<td>13.06</td>
</tr>
<tr>
<td>Max # of ICD9 codes per visit</td>
<td>41</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td># of category codes</td>
<td>157</td>
<td>155</td>
<td>171</td>
</tr>
<tr>
<td>Avg. # of category codes per visit</td>
<td>2.30</td>
<td>2.92</td>
<td>10.16</td>
</tr>
<tr>
<td>Max # of category codes per visit</td>
<td>23</td>
<td>22</td>
<td>30</td>
</tr>
</tbody>
</table>

- **Evaluation Measures**
  - Visit-level precision@$k$
  - Code-level accuracy@$k$
  - Vary $k$ from 5 to 30
Performance Evaluation

Results on three datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Visit-Level Precision@k</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Code-Level Accuracy@k</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>Medicaid</td>
<td>KAME</td>
<td>0.6107</td>
<td>0.7475</td>
<td>0.8168</td>
<td>0.8606</td>
<td>0.8920</td>
<td>0.9154</td>
<td>0.5461</td>
<td>0.7037</td>
<td>0.7808</td>
<td>0.8305</td>
<td>0.8667</td>
<td>0.8940</td>
</tr>
<tr>
<td></td>
<td>GRAM</td>
<td>0.5832</td>
<td>0.7189</td>
<td>0.7902</td>
<td>0.8367</td>
<td>0.8717</td>
<td>0.8976</td>
<td>0.5279</td>
<td>0.6842</td>
<td>0.7630</td>
<td>0.8146</td>
<td>0.8528</td>
<td>0.8819</td>
</tr>
<tr>
<td></td>
<td>Dipole</td>
<td>0.5943</td>
<td>0.7226</td>
<td>0.7892</td>
<td>0.8340</td>
<td>0.8680</td>
<td>0.8942</td>
<td>0.5406</td>
<td>0.6903</td>
<td>0.7637</td>
<td>0.8130</td>
<td>0.8503</td>
<td>0.8791</td>
</tr>
<tr>
<td></td>
<td>RNN+</td>
<td>0.5964</td>
<td>0.7210</td>
<td>0.7919</td>
<td>0.8397</td>
<td>0.8746</td>
<td>0.9011</td>
<td>0.5402</td>
<td>0.6867</td>
<td>0.7642</td>
<td>0.8166</td>
<td>0.8550</td>
<td>0.8845</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>0.5448</td>
<td>0.6737</td>
<td>0.7503</td>
<td>0.8036</td>
<td>0.8433</td>
<td>0.8740</td>
<td>0.4914</td>
<td>0.6370</td>
<td>0.7200</td>
<td>0.7782</td>
<td>0.8222</td>
<td>0.8564</td>
</tr>
<tr>
<td>Diabetes</td>
<td>KAME</td>
<td>0.5881</td>
<td>0.7313</td>
<td>0.8054</td>
<td>0.8523</td>
<td>0.8859</td>
<td>0.9107</td>
<td>0.5147</td>
<td>0.6939</td>
<td>0.7779</td>
<td>0.8293</td>
<td>0.8666</td>
<td>0.8949</td>
</tr>
<tr>
<td></td>
<td>GRAM</td>
<td>0.5596</td>
<td>0.7048</td>
<td>0.7822</td>
<td>0.8326</td>
<td>0.8684</td>
<td>0.8962</td>
<td>0.4958</td>
<td>0.6776</td>
<td>0.7617</td>
<td>0.8158</td>
<td>0.8546</td>
<td>0.8848</td>
</tr>
<tr>
<td></td>
<td>Dipole</td>
<td>0.5697</td>
<td>0.7015</td>
<td>0.7765</td>
<td>0.8267</td>
<td>0.8640</td>
<td>0.8921</td>
<td>0.5110</td>
<td>0.6771</td>
<td>0.7585</td>
<td>0.8120</td>
<td>0.8520</td>
<td>0.8824</td>
</tr>
<tr>
<td></td>
<td>RNN+</td>
<td>0.5680</td>
<td>0.7007</td>
<td>0.7769</td>
<td>0.8279</td>
<td>0.8649</td>
<td>0.8943</td>
<td>0.5086</td>
<td>0.6740</td>
<td>0.7569</td>
<td>0.8118</td>
<td>0.8519</td>
<td>0.8838</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>0.5515</td>
<td>0.6851</td>
<td>0.7639</td>
<td>0.8179</td>
<td>0.8575</td>
<td>0.8877</td>
<td>0.4984</td>
<td>0.6611</td>
<td>0.7459</td>
<td>0.8024</td>
<td>0.8445</td>
<td>0.8765</td>
</tr>
<tr>
<td>MIMIC-III</td>
<td>KAME</td>
<td>0.7103</td>
<td>0.6568</td>
<td>0.6967</td>
<td>0.7562</td>
<td>0.8091</td>
<td>0.8470</td>
<td>0.3167</td>
<td>0.5100</td>
<td>0.6379</td>
<td>0.7240</td>
<td>0.7862</td>
<td>0.8303</td>
</tr>
<tr>
<td></td>
<td>GRAM</td>
<td>0.6998</td>
<td>0.6447</td>
<td>0.6847</td>
<td>0.7439</td>
<td>0.8007</td>
<td>0.8424</td>
<td>0.3123</td>
<td>0.5026</td>
<td>0.6296</td>
<td>0.7142</td>
<td>0.7798</td>
<td>0.8266</td>
</tr>
<tr>
<td></td>
<td>Dipole</td>
<td>0.6220</td>
<td>0.5839</td>
<td>0.6310</td>
<td>0.6953</td>
<td>0.7556</td>
<td>0.8059</td>
<td>0.2774</td>
<td>0.4556</td>
<td>0.5801</td>
<td>0.6671</td>
<td>0.7354</td>
<td>0.7902</td>
</tr>
<tr>
<td></td>
<td>RNN+</td>
<td>0.6158</td>
<td>0.5803</td>
<td>0.6243</td>
<td>0.6912</td>
<td>0.7542</td>
<td>0.8017</td>
<td>0.2760</td>
<td>0.4548</td>
<td>0.5751</td>
<td>0.6647</td>
<td>0.7350</td>
<td>0.7867</td>
</tr>
<tr>
<td></td>
<td>RNN</td>
<td>0.6580</td>
<td>0.6186</td>
<td>0.6637</td>
<td>0.7254</td>
<td>0.7836</td>
<td>0.8272</td>
<td>0.2941</td>
<td>0.4836</td>
<td>0.6106</td>
<td>0.6961</td>
<td>0.7629</td>
<td>0.8119</td>
</tr>
</tbody>
</table>

- The performance of the proposed KAME is better than that of all the baselines on the three datasets.
- Fully utilizing medical knowledge graph is important!
- The proposed KAME achieves robust results on different datasets.
Data Sufficiency Evaluation

- We first divide medical codes into four groups: 0-25, 25-50, 50-75 and 75-100, based on their frequency in the training set.
- The 0-25 group represents the most rare codes in the training set, while codes in the 75-100 group are the most common ones.
- We then calculate the average accuracy of codes in each group on the testing set.
The benefit of KAME is to interpret the prediction results with the proposed knowledge-based attention mechanism.

The proposed KAME can learn an accurate attention weight for each piece of knowledge.

With the proposed knowledge attention mechanism, KAME can significantly improve the performance of the diagnosis prediction task in healthcare.
Interpretability Analysis

- **Interpretability of the learned medical code representations**
  - Randomly select 2000 medical codes and then plot on a 2-D space with t-SNE using their learned embeddings.

  ![Cluster Plot]

  - Each dot represents a diagnosis code. The colors of the dots represent the disease categories, i.e., *cluster labels*.
  - Ideally, the dots with the *same color* should be in the *same cluster*, and there are *margins* among *different clusters*. 
Conclusions

- We propose KAME, an end-to-end, accurate and robust model to accurately predict patients' future visit information with medical ontologies, which explicitly makes use of medical knowledge in the whole prediction process.

- We design a novel knowledge-level attention mechanism, which significantly helps the proposed KAME to improve the predictive performance.

- We empirically show that the proposed KAME has strong robustness and outperforms existing methods in diagnosis prediction on three real world datasets.

- We qualitatively demonstrate the interpretability of the learned representations of medical codes and qualitatively validate the reasonableness of the designed knowledge attention mechanism.
Travel Award

Thank You!

Questions?

More studies in Healthcare can be found at http://www.acsu.buffalo.edu/~fenglong.