The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining August 14–18, 2021, Singapore.

RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation

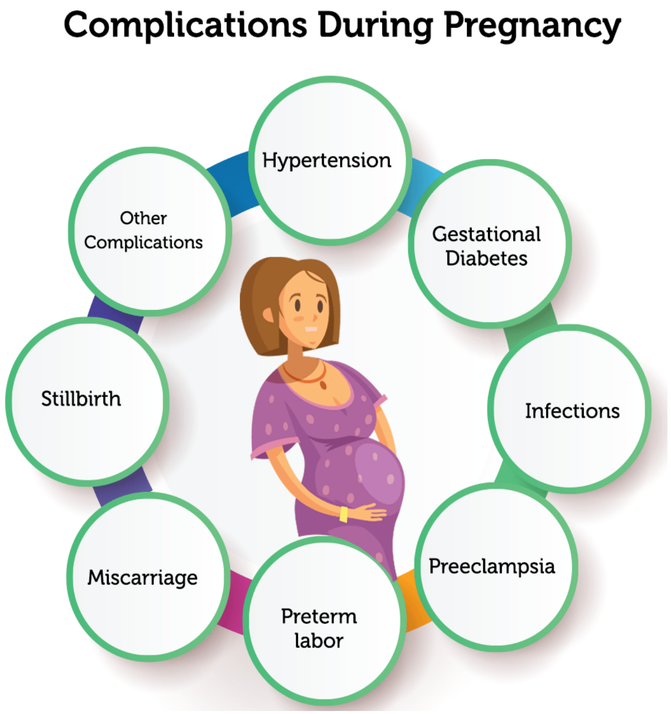
Houxing Ren, Jingyuan Wang, Wayne Xin Zhao, Ning Wu

Beihang University, Beijing, China

Renmin University of China, Beijing, China

Background

- Pregnancy complications, such as gestational diabetes and gestational hypertension, create severe threats to the health of pregnant women.
- It has been reported that about **300,000 women** died due to complications in pregnancy and childbirth in 2017.



Motivation

Early Study

- For specific task.
  - Diagnosis prediction
  - Risk prediction
- To address unique issues.
  - Irregular time intervals
  - Data insufficiency

It is difficult to reuse these existing methods to provide a general solution for pregnancy complications!

How to learn effective representations from EHR data, which can capture the major data characteristics of examination records?

Challenges

- EHR data change with irregular time intervals.
  - examination records of prenatal care correspond to **irregularly distributed samples** of women's physical characteristics during the entire pregnancy

- Different pregnancy complications usually correspond to varying factors or indicators.
  - gestational diabetes is more sensitive to **timesteps**
  - gestational hypertension is more sensitive to **specific week**

- The EHR data tend to be sparse or incomplete.
  - only **a few items** are checked at each visit

Idea

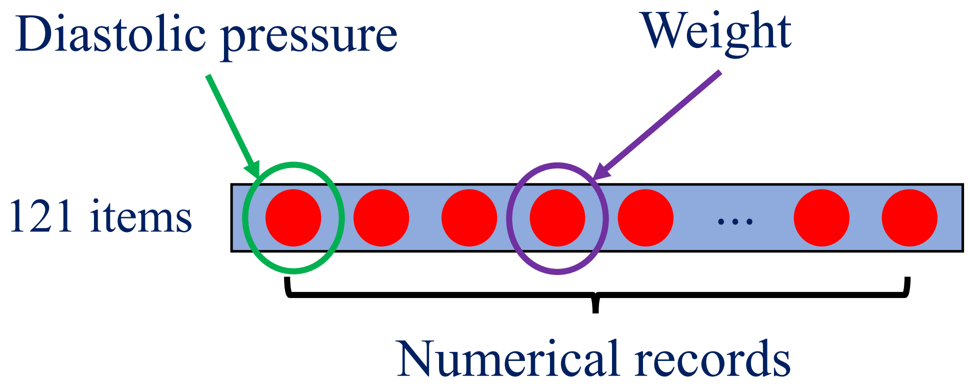
- Learn robust representation with **pre-training technique.**

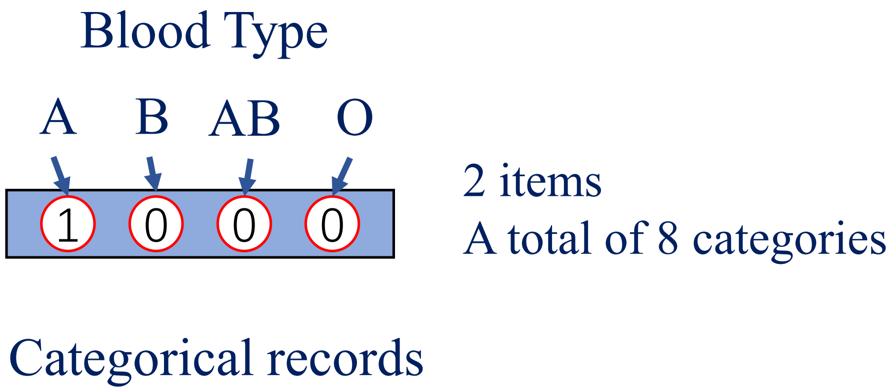
- Design a suitable network architecture for pretraining on EHR data.

- Design pre-training tasks that can effectively extract data characteristics and address EHR data issues.

Definition

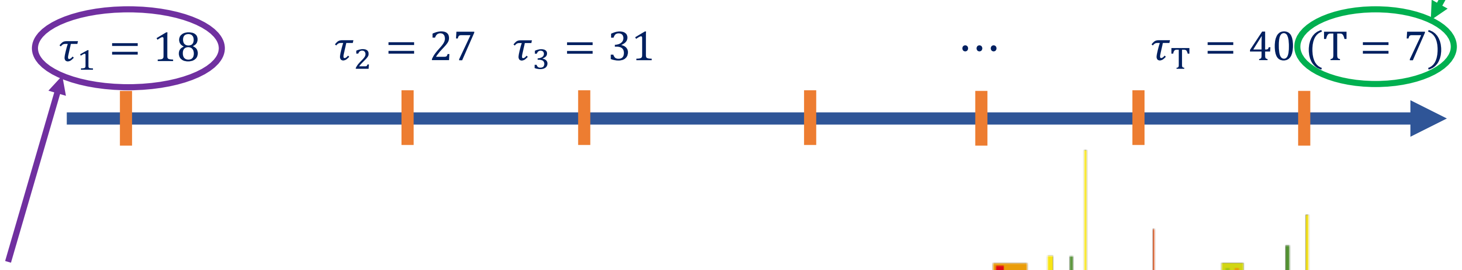
■ Examination Record.





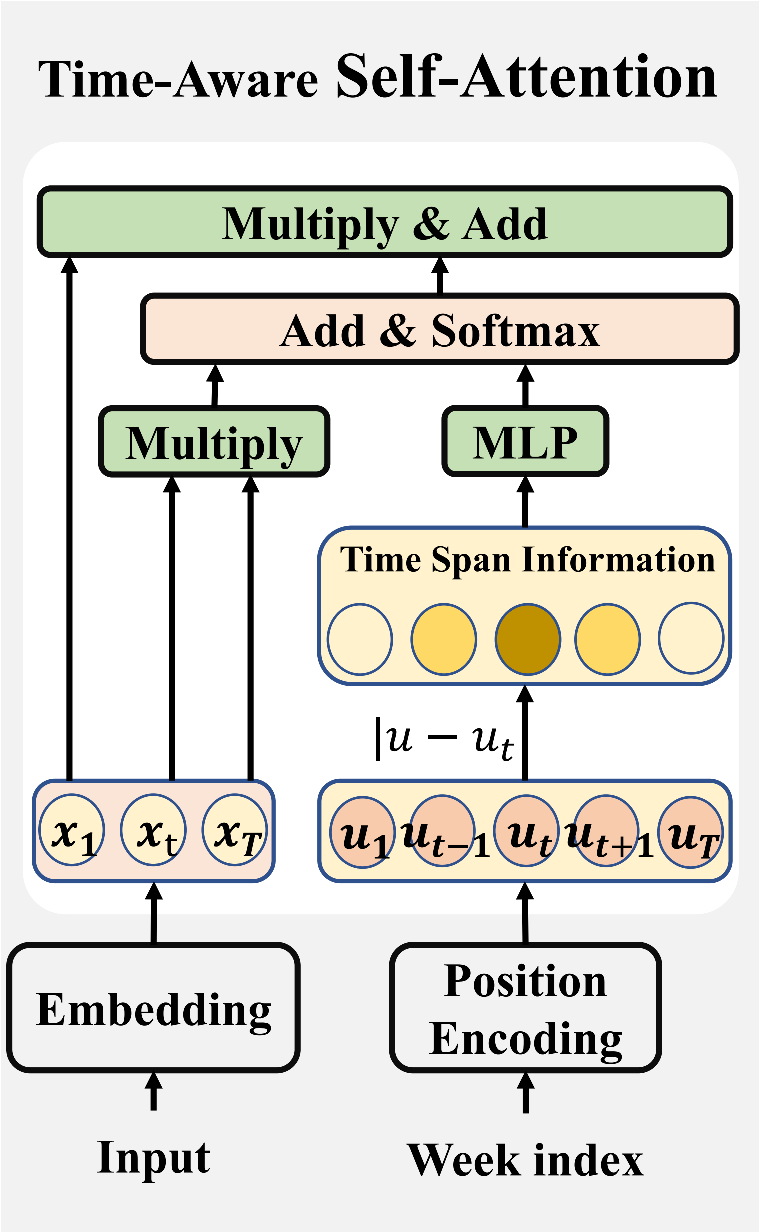
■ Prenatal Care Sequence Data.

Different T for different pregnant women

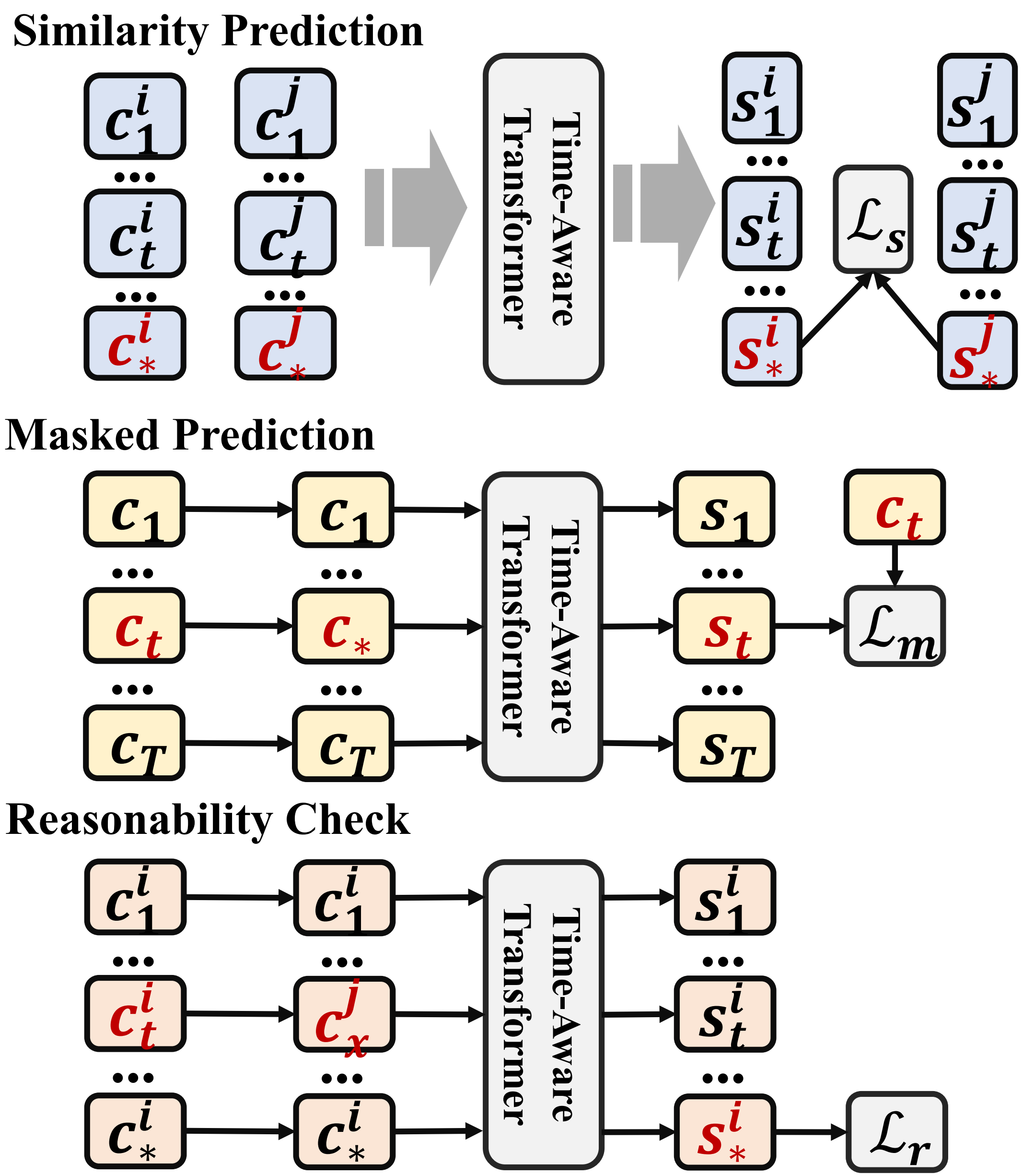


Different τ_i for different pregnant women

Overview

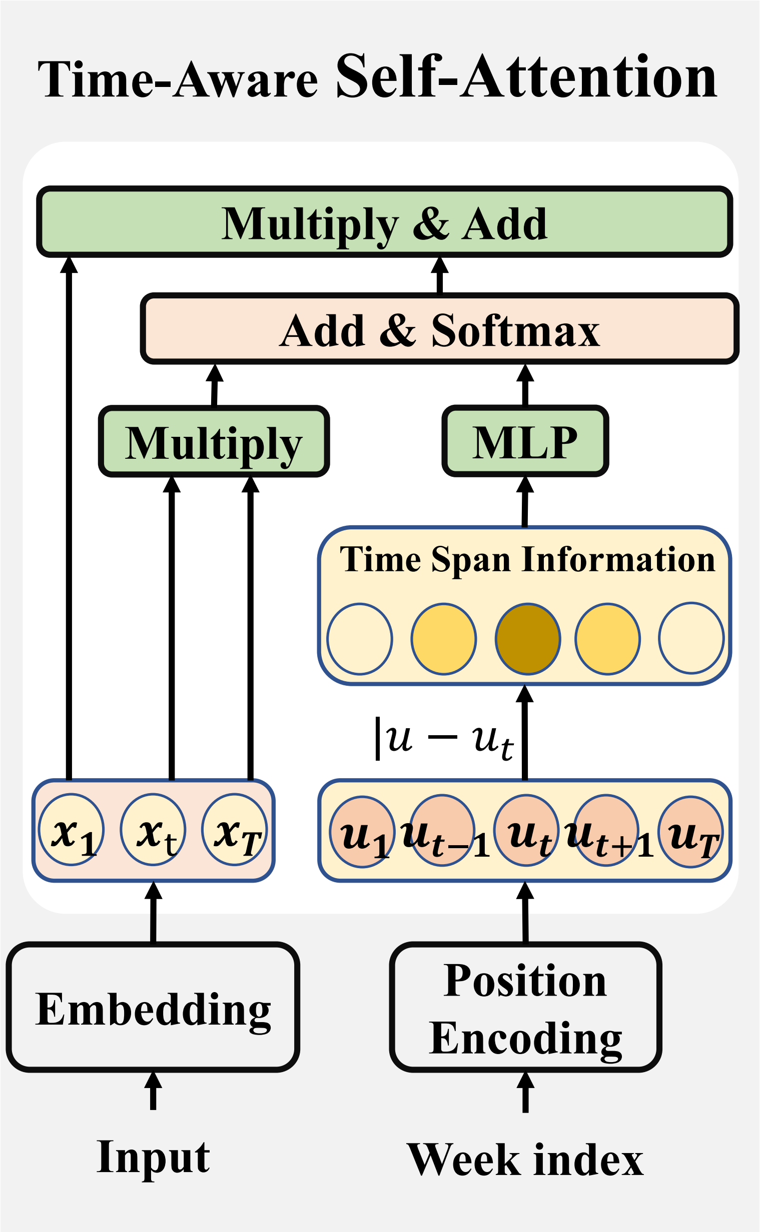


Time-Aware Self-Attention
Multiply & Add
Add & Softmax
Multiply
MLP
Time Span Information
$|u - u_t|$
x_1 x_t x_T
u_1 u_{t-1} u_t u_{t+1} u_T
Embedding
Position Encoding
Input
Week index



Similarity Prediction
Time-Aware Transformer
$\mathcal{L}_s$
Masked Prediction
Time-Aware Transformer
$\mathcal{L}_m$
Reasonability Check
Time-Aware Transformer
$\mathcal{L}_r$

Time-Aware Self-Attention



- Standard Self-Attention

$$A_{i,j} = \frac{q_i^T k_j}{\sqrt{h}}$$

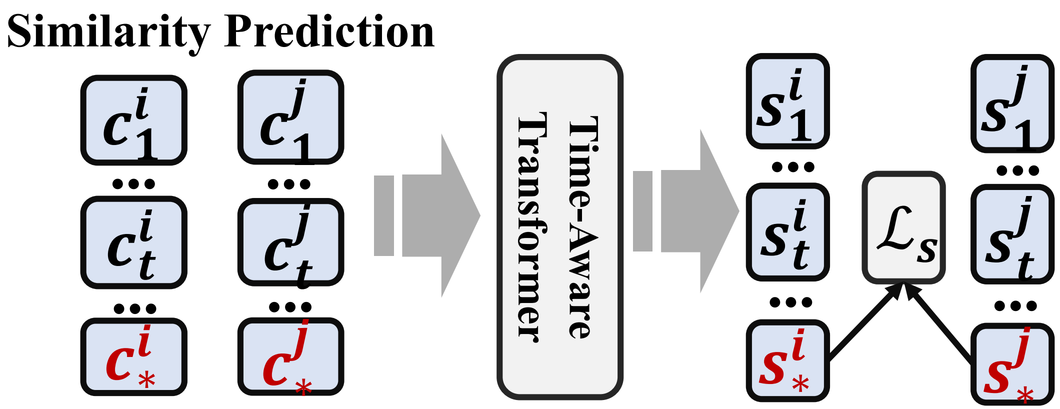
- Self-Attention with week index

$$A_{i,j} = \frac{x_i^T x_j + x_i^T u_j + u_i^T x_j + u_i^T u_j}{\sqrt{h}}$$

- Time-Aware Self-Attention

$$A_{i,j} = \frac{q_i^T k_j + \boldsymbol{W_\tau} \times |\boldsymbol{u_i} - \boldsymbol{u_j}|}{\sqrt{h}}$$

Similarity Prediction

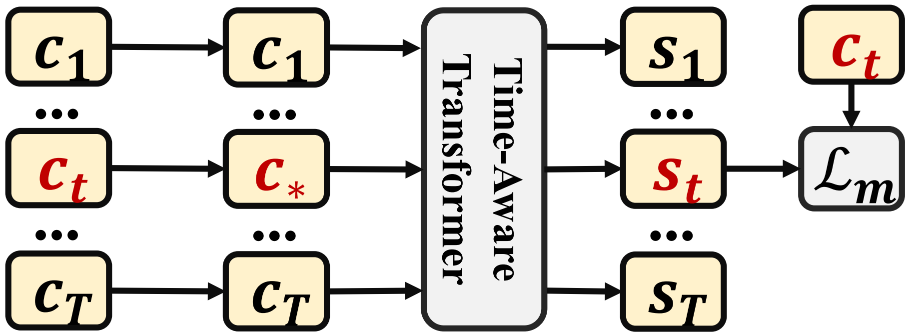


- Measure the Euclidean distance of all pregnant women's last visits.
- Take the 15% pairs with the smallest distance as the positive samples and the 15% pairs with the largest distance as the negative samples to train the model.

$$\mathcal{L}_s = \frac{1}{N_p}\sum_{i=1}^{N_p} z_i d_i^2 + (1 - z_i)max(m - d_i, 0)^2$$

Masked Prediction

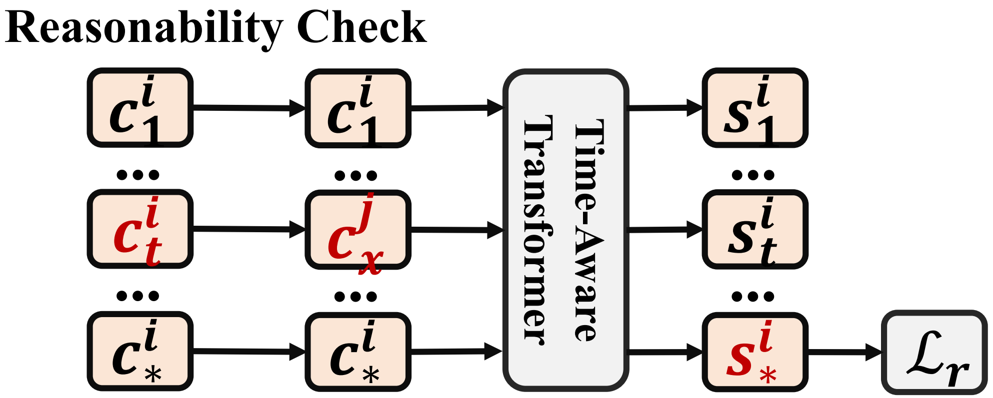
Masked Prediction



- Randomly mask 30% of all visits by c_*.
- Use the corresponding hidden state to predict the important examination records.

$$\mathcal{L}_{\mathrm{m}} = \frac{1}{|C^{\dagger}|} \sum_{c^{\dagger} \in C^{\dagger}} ||\hat{c}^{\dagger} - c^{\dagger}||_2^2$$

Reasonability Check



- Negative samples: randomly select 50% - 75% visits and replace them with visits from other sequences.
- Positive samples: do nothing.

$$\mathcal{L}_{\mathrm{r}} = -\frac{\mathbf{1}}{\boldsymbol{N}_{\mathrm{r}}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}_{\mathrm{r}}} \left(\mathrm{r}_i \log(\hat{r}_i) + (\mathbf{1} - \mathrm{r}_i) \log(1 - \hat{r}_i) \right)$$

Model Training

- Pre-training.
  - Pre-train the model with three pre-training tasks for robust representation.
- Fine-tuning.
  - Fine-tune the model with specific task for better performance.

Algorithm 1 The training algorithm for the RAPT model.

Input: A pregnant women examination records dataset X.
Output: Model parameters θ_I, θ_T and θ_F.

Randomly initialize $\theta_I, \theta_T, \theta_P$ and θ_F.
for episode = 1 to epoch **do**
    Calculate representations of visit sequence $\boldsymbol{S}^{(i)}$ by Eq. (8).
    Calculate loss of pre-training by Eq. (14).
    Perform stochastic gradient descent on Eq. (14) *w.r.t.* θ_I, θ_T and θ_P.
end for
Drop θ_P and keep other parameters.
for episode = 1 to epoch **do**
    Calculate representations of visit sequence $\boldsymbol{s}_*^{(i)}$ by Eq. (8).
    Calculate loss of fine-tuning by Eq. (16) or Eq. (18).
    Perform stochastic gradient ascent on Eq. (16) or Eq. (18) *w.r.t.* θ_I, θ_T and θ_F.
end for
return θ_I, θ_T and θ_F.

Dataset & Task

- **Dataset.**
  - A hospital in Beijing
  - From 2008 to 2018
  - 63,001 pregnant women

Dataset statistics

Dataset	Pre-train	Diab.	Hype.	Outcome	Period
# of samples	63,001	20,160	5,744	8,514	1,556
# of visits	427,369	137,873	38,600	57,081	19,434
Avg. # of visits	6.78	6.84	6.72	6.70	12.49
Avg. week of FV	13.82	14.46	14.51	14.50	14.63
Avg. week of LV	28.18	28.23	28.21	28.20	36.96

- **Downstream Task**
  - Gestational Diabetes Prediction
  - Gestational Hypertension Prediction
  - Pregnancy Outcome Prediction.
  - Risk Period Prediction.

Metric & Baseline

Metric

- Classification Task： AUC, Precision, Recall, F1, ACC.
- Regression Task： RMSE, MAE, MPAE, R2, EV.

Baseline

- LSTM [Neural Comput. 1997] : Long Short-Term Memory.
- Transformer [NIPS 2017] : based solely on attention mechanisms.
- RETAIN [NIPS 2016] : Reverse Time Attention.
- T-LSTM [KDD 2017] : Time-aware LSTM.
- Dipole [KDD 2017] : Diagnosis prediction model.
- HiTANet [KDD 2020] : Hierarchical Time-aware Attention Network.

Results & Analysis

Task		Diabetes Prediction					Task		Hypertension Prediction				
Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	0.670	0.559	0.934	0.699	0.738	Model	LSTM	0.735	0.703	0.775	0.743	0.810
	Trans.	0.737	0.643	0.872	0.740	0.811		Trans.	0.733	0.677	0.826	0.744	0.800
	RETAIN	0.644	0.522	**0.971**	0.679	0.708		RETAIN	0.738	0.681	0.812	0.741	0.814
	T-LSTM	0.726	0.631	0.891	0.739	0.795		T-LSTM	0.738	0.625	**0.901**	0.738	0.815
	Dipole	0.724	0.675	0.794	0.730	0.790		Dipole	0.737	**0.730**	0.746	0.738	0.812
	HiTANet	0.747	0.723	0.764	0.743	0.813		HiTANet	0.739	0.718	0.777	0.746	0.811
	RAPT	**0.807**	**0.836**	0.763	**0.798**	**0.867**		RAPT	**0.746**	0.671	0.840	**0.749**	**0.820**

Task		Pregnancy Outcome Prediction					Task		Risk Period Prediction				
Metric		RMSE ↓	MAE ↓	MAPE ↓	R2 ↑	EV ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	10.661	7.449	0.094	0.000	0.000	Model	LSTM	0.909	0.770	0.838	0.802	0.959
	Trans.	8.620	5.319	0.068	0.338	0.339		Trans.	0.908	0.767	0.808	0.784	0.947
	RETAIN	9.046	5.812	0.081	0.246	0.261		RETAIN	0.848	0.550	0.694	0.613	0.854
	T-LSTM	10.664	7.454	0.104	-0.001	0.000		T-LSTM	0.908	0.772	0.821	0.795	0.960
	Dipole	9.229	6.200	0.079	0.232	0.233		Dipole	0.918	0.807	0.824	0.812	0.965
	HiTANet	8.631	5.377	0.077	0.337	0.337		HiTANet	0.900	0.759	0.775	0.767	0.943
	RAPT	**8.525**	**5.184**	**0.063**	**0.350**	**0.352**		RAPT	**0.976**	**0.964**	**0.925**	**0.944**	**0.985**

- Gestational diabetes is more sensitive to **timesteps**.
- Models considering **irregular time intervals** achieve better performance.

Results & Analysis

Task		Diabetes Prediction					Task		Hypertension Prediction				
Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	0.670	0.559	0.934	0.699	0.738	Model	LSTM	0.735	0.703	0.775	0.743	0.810
	Trans.	0.737	0.643	0.872	0.740	0.811		Trans.	0.733	0.677	0.826	0.744	0.800
	RETAIN	0.644	0.522	**0.971**	0.679	0.708		RETAIN	0.738	0.681	0.812	0.741	0.814
	T-LSTM	0.726	0.631	0.891	0.739	0.795		T-LSTM	0.738	0.625	**0.901**	0.738	0.815
	Dipole	0.724	0.675	0.794	0.730	0.790		Dipole	0.737	**0.730**	0.746	0.738	0.812
	HiTANet	0.747	0.723	0.764	0.743	0.813		HiTANet	0.739	0.718	0.777	0.746	0.811
	RAPT	**0.807**	**0.836**	0.763	**0.798**	**0.867**		RAPT	**0.746**	0.671	0.840	**0.749**	**0.820**
Task		Pregnancy Outcome Prediction					Task		Risk Period Prediction				
Metric		RMSE ↓	MAE ↓	MAPE ↓	R2 ↑	EV ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	10.661	7.449	0.094	0.000	0.000	Model	LSTM	0.909	0.770	0.838	0.802	0.959
	Trans.	8.620	5.319	0.068	0.338	0.339		Trans.	0.908	0.767	0.808	0.784	0.947
	RETAIN	9.046	5.812	0.081	0.246	0.261		RETAIN	0.848	0.550	0.694	0.613	0.854
	T-LSTM	10.664	7.454	0.104	-0.001	0.000		T-LSTM	0.908	0.772	0.821	0.795	0.960
	Dipole	9.229	6.200	0.079	0.232	0.233		Dipole	0.918	0.807	0.824	0.812	0.965
	HiTANet	8.631	5.377	0.077	0.337	0.337		HiTANet	0.900	0.759	0.775	0.767	0.943
	RAPT	**8.525**	**5.184**	**0.063**	**0.350**	**0.352**		RAPT	**0.976**	**0.964**	**0.925**	**0.944**	**0.985**

- Gestational hypertension is more sensitive to examination records of **specific weeks**.
- Models considering **irregular time intervals** perform worse than models considering **other characteristics**.

Results & Analysis

Task		Diabetes Prediction					Task		Hypertension Prediction				
Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	0.670	0.559	0.934	0.699	0.738	Model	LSTM	0.735	0.703	0.775	0.743	0.810
	Trans.	0.737	0.643	0.872	0.740	0.811		Trans.	0.733	0.677	0.826	0.744	0.800
	RETAIN	0.644	0.522	**0.971**	0.679	0.708		RETAIN	0.738	0.681	0.812	0.741	0.814
	T-LSTM	0.726	0.631	0.891	0.739	0.795		T-LSTM	0.738	0.625	**0.901**	0.738	0.815
	Dipole	0.724	0.675	0.794	0.730	0.790		Dipole	0.737	**0.730**	0.746	0.738	0.812
	HiTANet	0.747	0.723	0.764	0.743	0.813		HiTANet	0.739	0.718	0.777	0.746	0.811
	RAPT	**0.807**	**0.836**	0.763	**0.798**	**0.867**		RAPT	**0.746**	0.671	0.840	**0.749**	**0.820**
Task		Pregnancy Outcome Prediction					Task		Risk Period Prediction				
Metric		RMSE ↓	MAE ↓	MAPE ↓	R2 ↑	EV ↑	Metric		ACC ↑	Pre ↑	Recall ↑	F1 ↑	AUC ↑
Model	LSTM	10.661	7.449	0.094	0.000	0.000	Model	LSTM	0.909	0.770	0.838	0.802	0.959
	Trans.	8.620	5.319	0.068	0.338	0.339		Trans.	0.908	0.767	0.808	0.784	0.947
	RETAIN	9.046	5.812	0.081	0.246	0.261		RETAIN	0.848	0.550	0.694	0.613	0.854
	T-LSTM	10.664	7.454	0.104	-0.001	0.000		T-LSTM	0.908	0.772	0.821	0.795	0.960
	Dipole	9.229	6.200	0.079	0.232	0.233		Dipole	0.918	0.807	0.824	0.812	0.965
	HiTANet	8.631	5.377	0.077	0.337	0.337		HiTANet	0.900	0.759	0.775	0.767	0.943
	RAPT	**8.525**	**5.184**	**0.063**	**0.350**	**0.352**		RAPT	**0.976**	**0.964**	**0.925**	**0.944**	**0.985**

- Our model handles **various characteristics** in EHR data.
- Our model is consistently better than all of the baselines in all tasks.

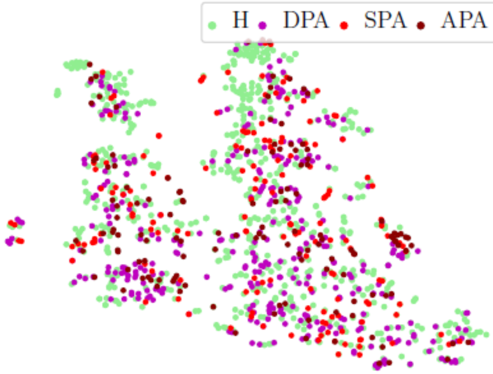
Results & Analysis

Metric	ACC	Precision	Recall	F1	AUC
Human	**0.763**	**1.000**	0.540	0.701	0.770
RAPT	0.746	0.671	**0.840**	**0.749**	**0.820**

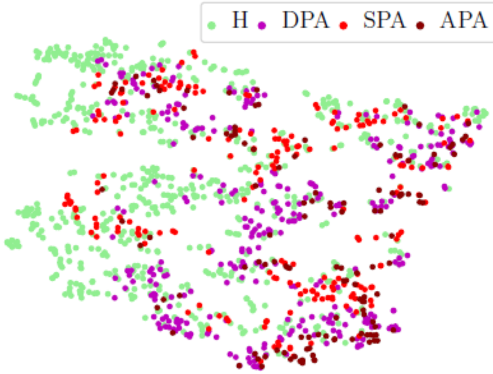
- Human performance is measured with the gold standard.
- The gold standard do not have the ability to **predict future examination records**.
- The diagnosis of our model is **timelier**.

Qualitative Analysis

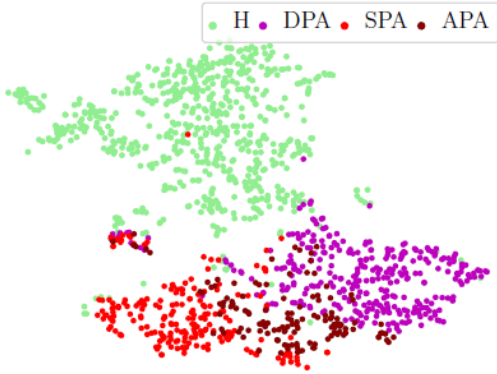
Scatter plots for embeddings



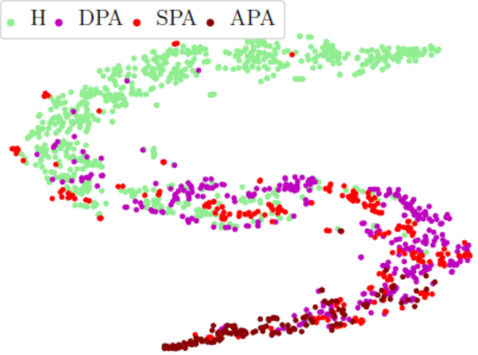
(a) Without training ($S = -0.08$).



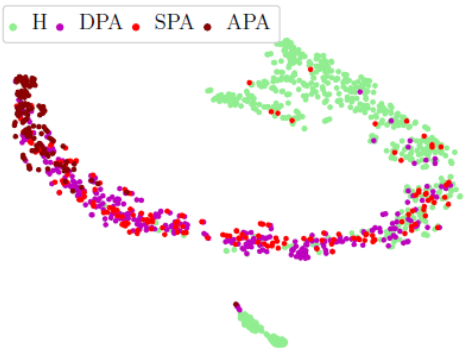
(b) Pre-trained ($S = -0.02$).



(c) Fine-tuned ($S = 0.33$).

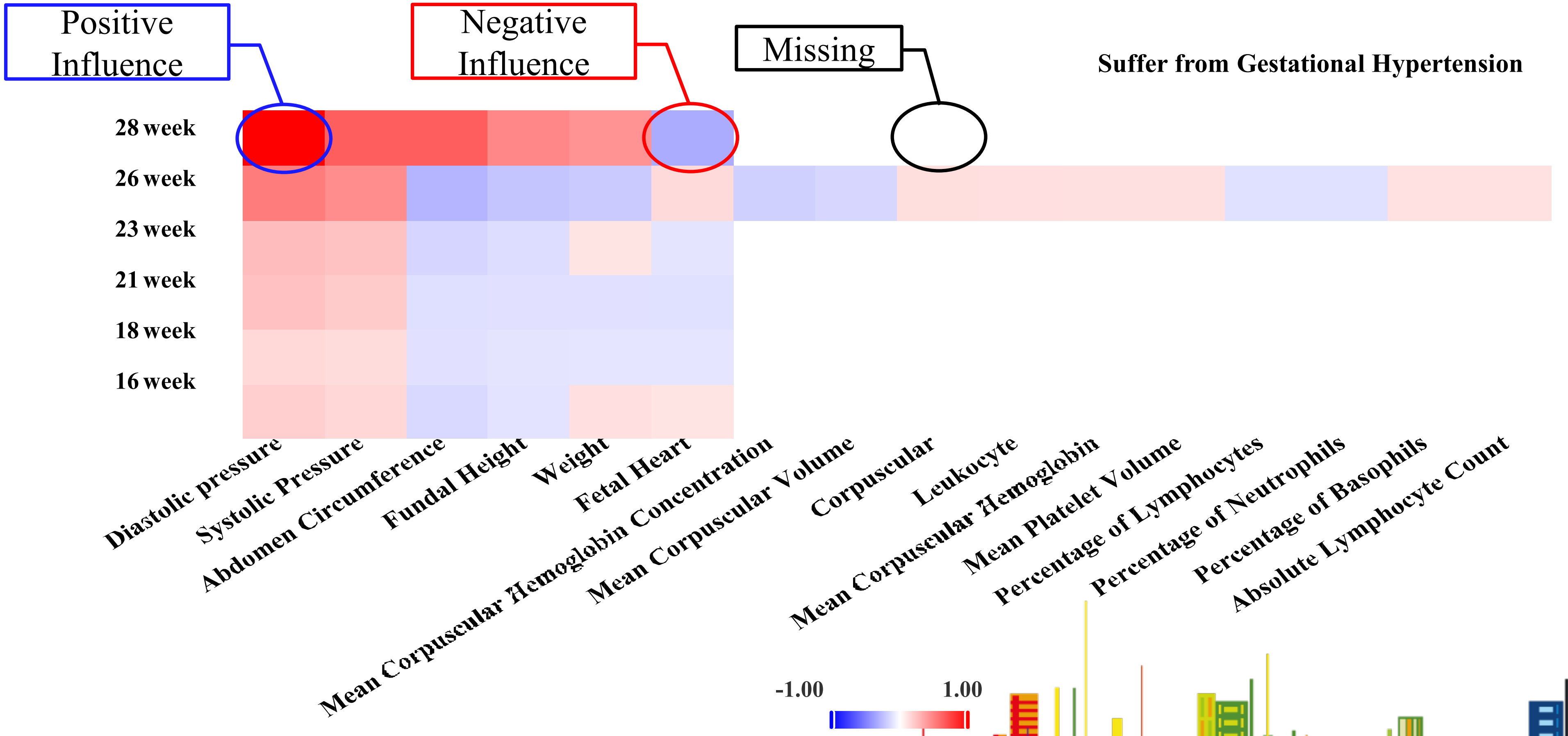


(d) Without pre-training ($S = 0.16$).



(e) Dipole ($S = 0.12$).

Interface for doctors



Conclusion

- We design a novel network architecture which is suitable for modeling EHR data and pre-training.

- We carefully design three pre-training tasks for medical data related to pregnancy complications.

- We introduced an interpretation method by sensitivity analysis and designed an interface to show the prediction results and interpretation.

Thank You!