Imperial College London

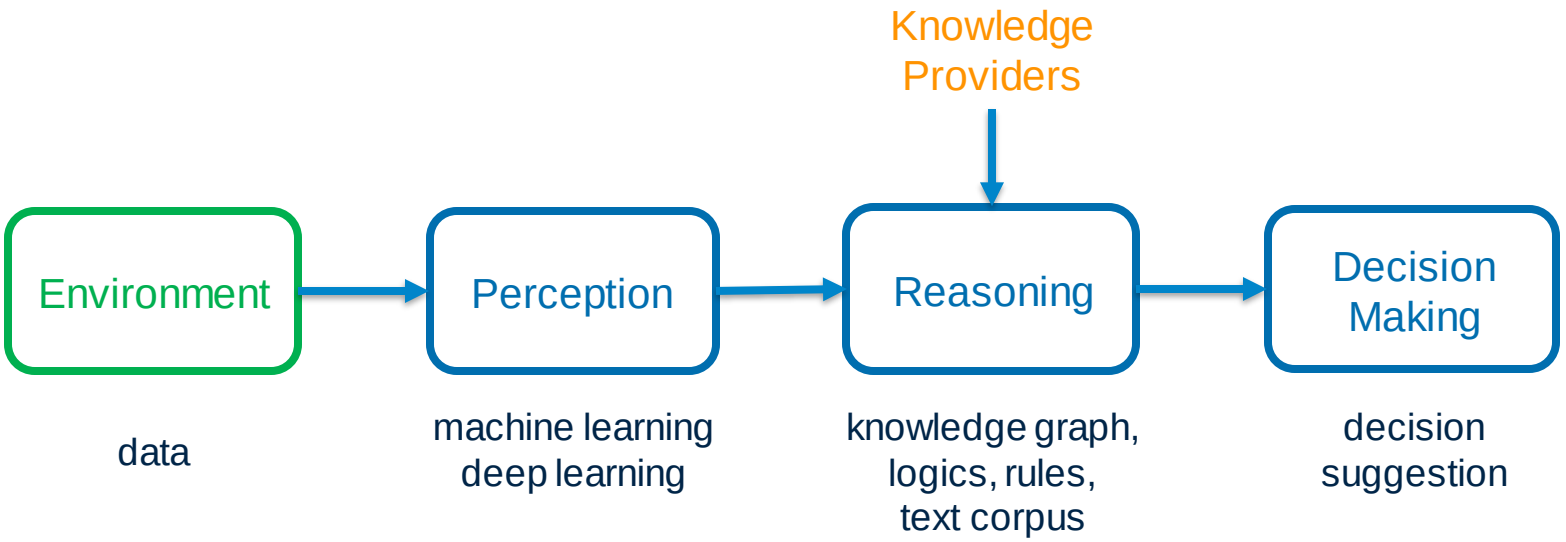
Integrating Prior Knowledge with Learning in Natural Language Processing

Unsupervised Phenotype Annotation via Semantic Latent Representations on Electronic Health Records

Jingqing Zhang
Data Science Institute
Imperial College London

Imperial College London

Deep Learning with Prior Knowledge



Imperial College London

Prior Knowledge

- Unstructured knowledge
  - Which can be implicitly contained in large text corpus.
  - Great success being used together with pre-training techniques.
- Structured knowledge
  - Which can be explicitly defined by knowledge graph, ontologies.
  - Proven to be effective in tasks that require reasoning and understanding.

Imperial College London

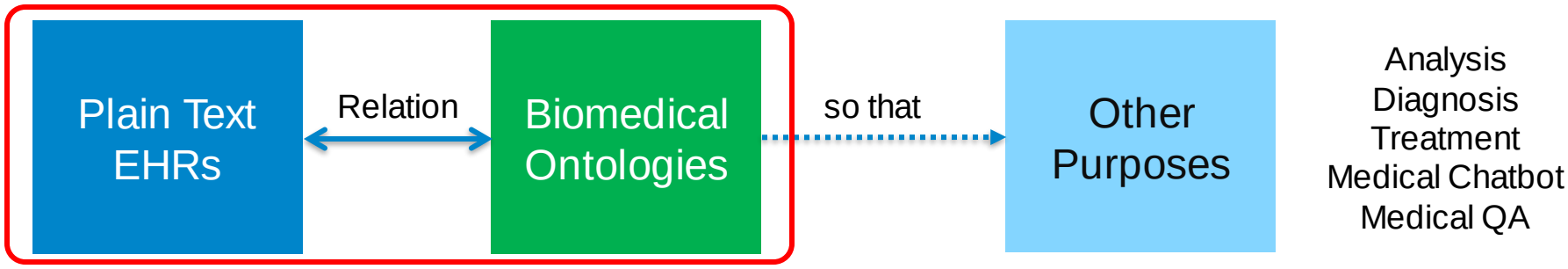
Our Research at Imperial DSI

- Leveraging structured and unstructured data as prior knowledge to improve deep learning models in natural language processing.
- Using structured data
  - Integrating Semantic Knowledge to Tackle Zero-shot Text Classification. NAACL 2019.
  - Unsupervised Annotation of Phenotypic Abnormalities via Semantic Latent Representations on Electronic Health Records. IEEE BIBM 2019.
- Using unstructured data
  - PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Submitted to ICML 2020.

Imperial College London

Motivation

- Electronic health records (EHRs) are widely and increasingly adopted.
- Understanding the professional and natural languages in medical domain with manually pre-defined biomedical ontologies.

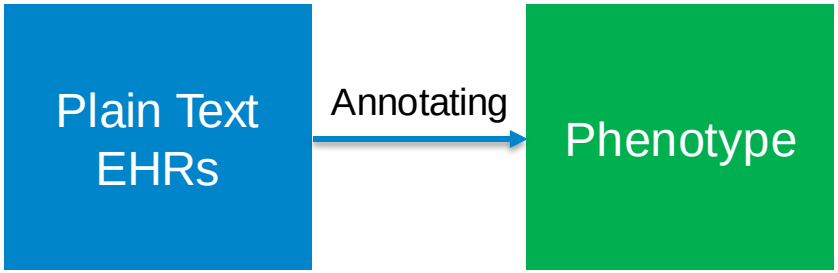


- International Classification of Diseases
- Human Phenotype Ontology
- Medical Subject Headings
- Online Mendelian Inheritance in Man
- …

Imperial College London

Motivation

- We aim to annotate EHRs with pre-defined phenotypes.
- EHRs serve as a rich source of phenotype information.
- We find the patients who were diagnosed as the same disease could be further classified into sub-groups by phenotypes.
- Phenotype annotation on EHRs can help disease diagnosis, and genomic diagnostic, towards precision medicine.



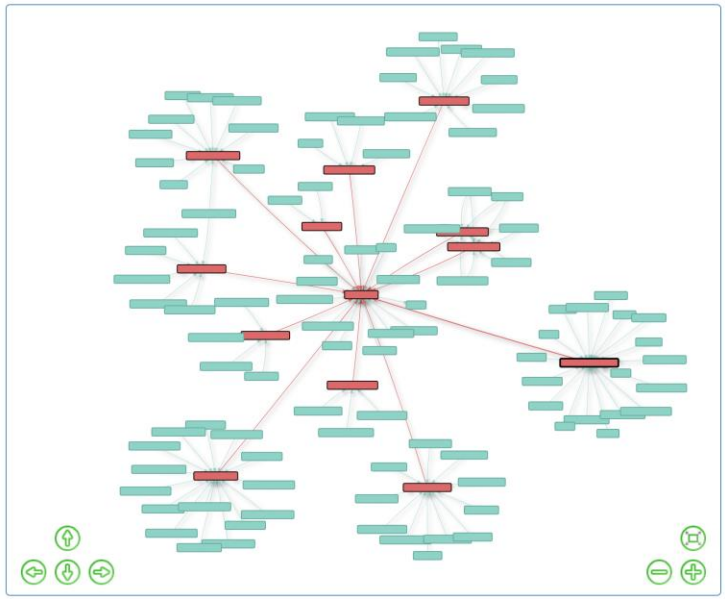
- Human Phenotype Ontology (HPO)
- Phenotype: observable characteristics
- Examples:
  - Abnormality of the digestive system
  - Abnormality of the immune system
  - Abnormality of the nervous system

Imperial College London

Datasets

- MIMIC-III
- A public EHRs database with 52k notes from 40k patients

- Human Phenotype Ontology (HPO)
- Standardize 13k HPO terms and each with description.



Imperial College London

Previous Works

- Information retrieval based approaches
  - E.g. OBO Annotator, NCBO Annotator, Bio-LarK, MetaMap, etc.
  - Suffers from computational inefficiency.
- Deep learning models
  - E.g. CNN [1]
  - Requires gold standards which is hard and expensive to acquire.

[1] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt,J. Foote Jr, E. T. Moseley, D. W. Grant, P. D.Tyleret al., "Comparing deep learning and concept extraction based methods for patient pheno-typing from clinical narratives,"PloS one, vol. 13, no. 2, p. e0192360,2018.

Imperial College London

Our Work

- We propose a novel unsupervised deep learning framework to exploit supportive phenotype knowledge in HPO and annotate general phenotypes from EHRs semantically.
- We demonstrate that our proposed method achieves state-of-the-art annotation performance and computational efficiency compared with other methods.

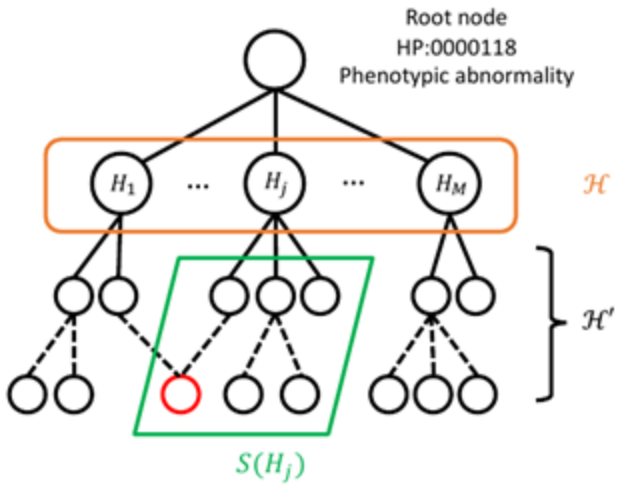
Imperial College London

Problem Formulation (Example)

EHRs snippets	Potential annotation with HPO terms
"Your blood pressure medications were adjusted to better control your blood pressure while you were in DKA"	• Abnormality of the cardiovascular system • Abnormality of the digestive system
"Given your complaints of chronic cough and heartburn, you should also discuss beginning a trial of a proton pump inhibitor such as Nexium or Prilosec to see if this helps your symptoms"	• Abnormality of the respiratory system • Abnormality of the digestive system

Problem Formulation

- There are two types of data sources.
  - $\mathcal{X} = \{X_1, \ldots, X_N\}$: a collection of EHRs and each EHR consists of textual notes written by clinicians.
  - $\mathcal{H} = \{H_1, \ldots, H_M\}$: a standardized general category of human phenotypes provided by HPO.
  - The HPO also provides additional subclasses $\mathcal{H}'$



Problem Formulation

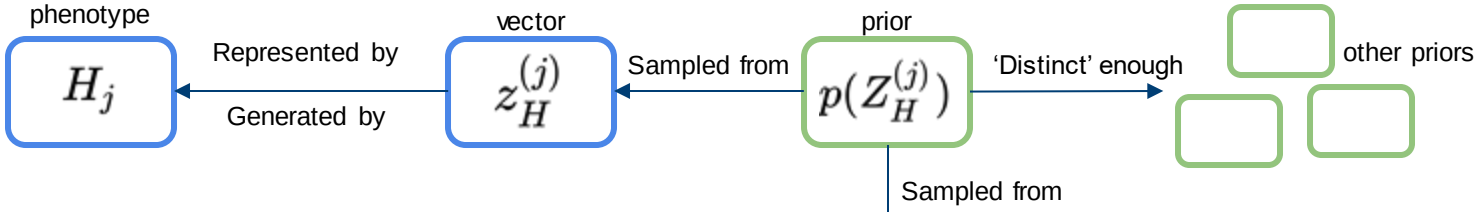
- The EHR can include either multiple phenotypes or a single or none. Therefore, learning the annotation of phenotypes from EHRs is essentially learning the conditional probability:

$$p(1_{H_j} | X_i)$$

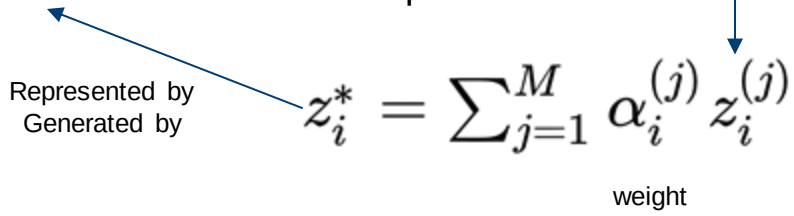
- I.e. a binary classification for each H_j
- As a whole, a multi-label classification on $\mathcal{H}$

Semantic Latent Representations

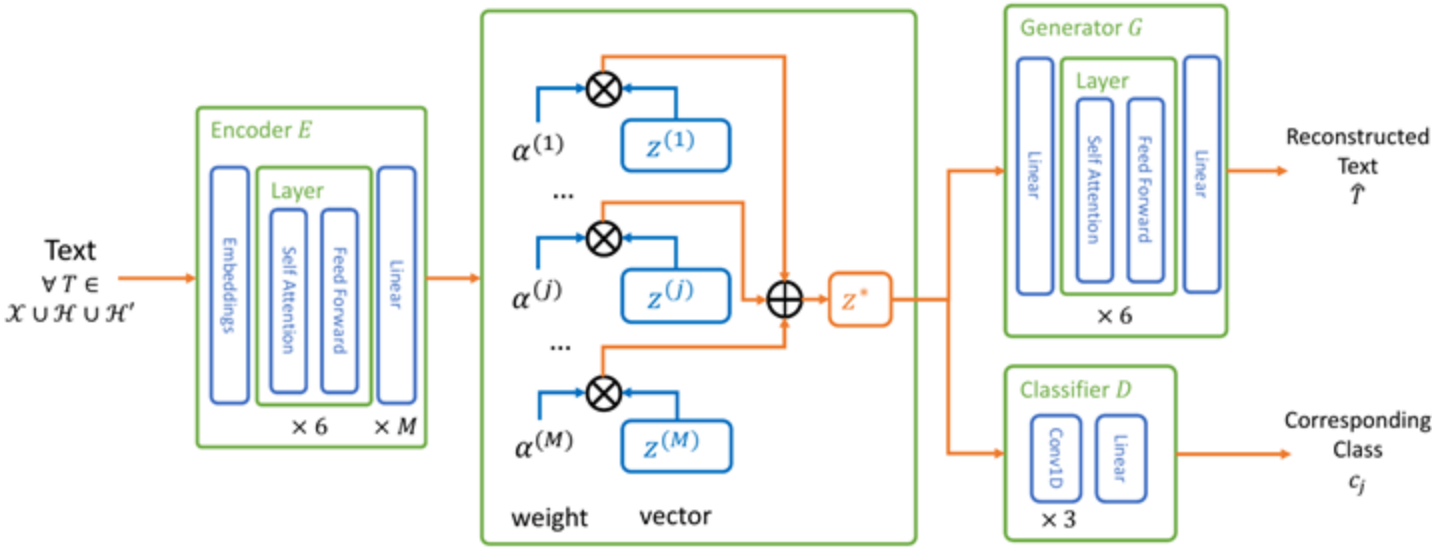
- Assumptions
  - The semantics of a general phenotype is represented by a prior distribution. The prior distribution of each phenotype should be 'distinct' enough from each other.



  - The semantics of EHR is a composition of the semantics of phenotypes.



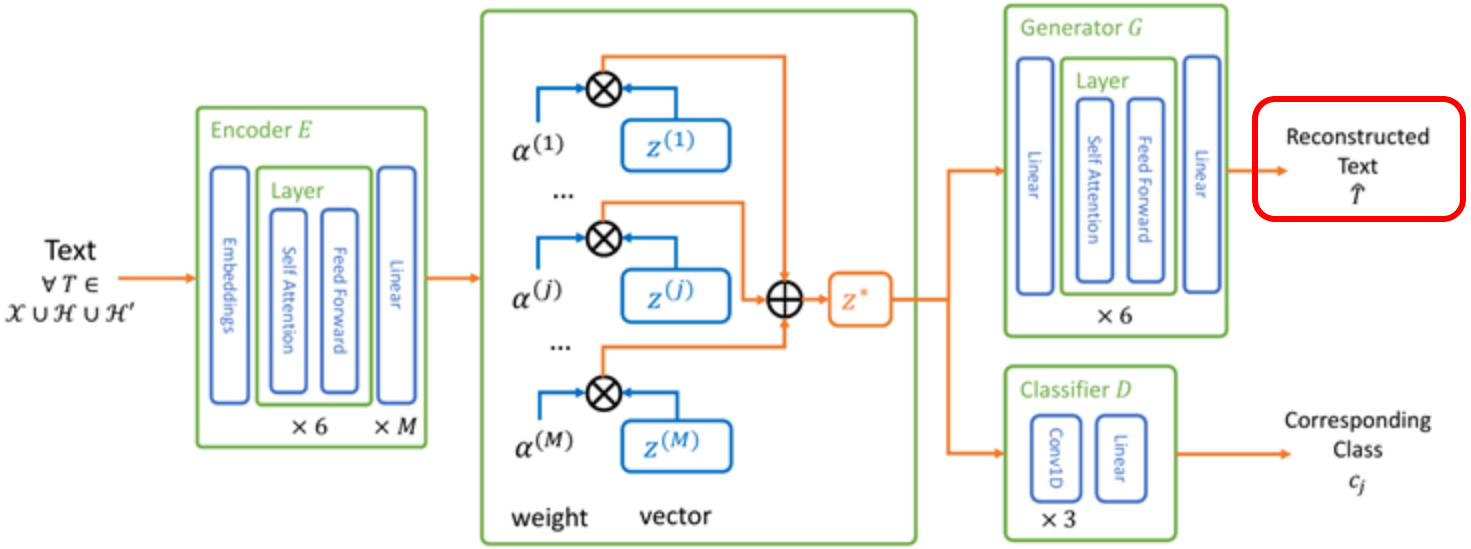
An Auto-Encoder Model



- General text reconstruction

$$\max \mathbb{E}_{Z \sim p(z^*|T)} \left[p(T|Z) \right] = \max_{\theta_E, \theta_G} \mathbb{E}_{Z \sim p_E(T;\theta_E)} \left[p_G(Z;\theta_G) \right]$$

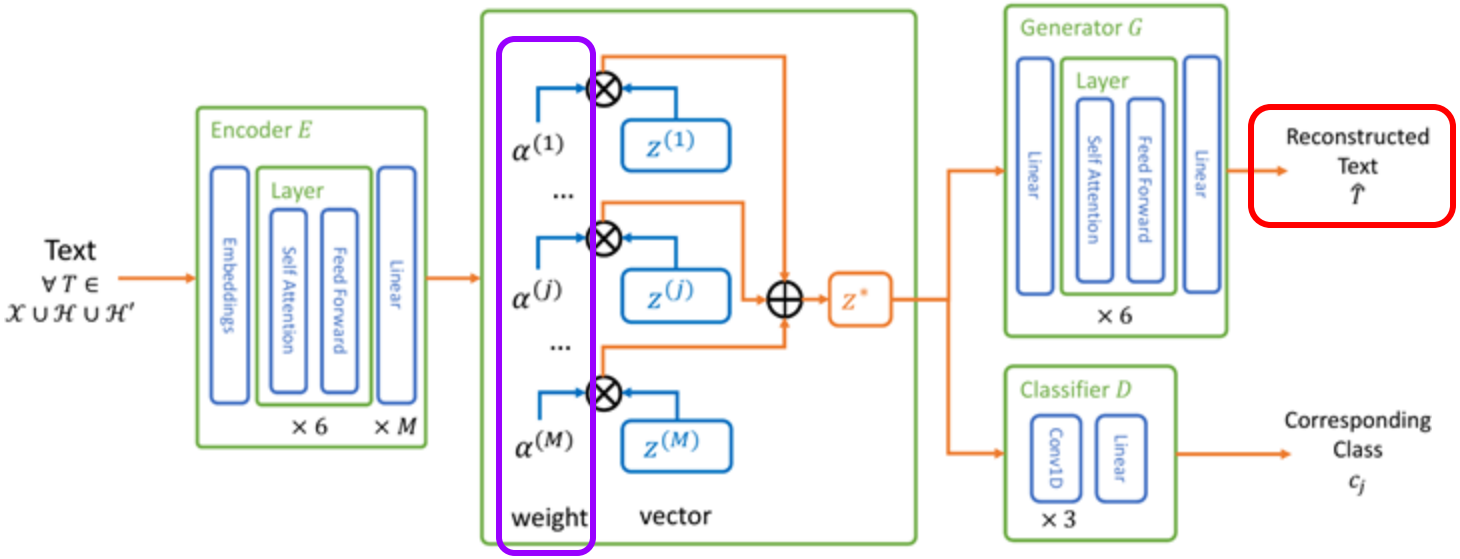
An Auto-Encoder Model



- Loss 1: text reconstruction of EHRs.

$$\mathcal{L}_{\text{rec}}^{X} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log p_G(X_i|E(X_i))\right]$$

An Auto-Encoder Model

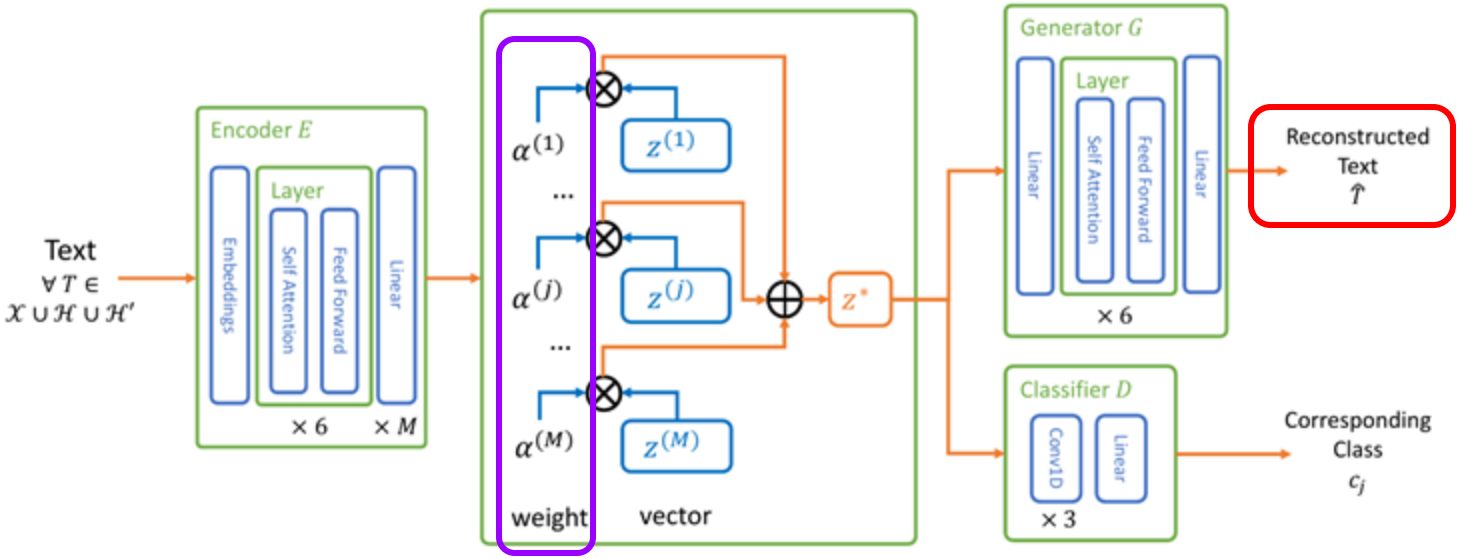


- Loss 2: text reconstruction of the general phenotypes' description.

$$\mathcal{L}_{\text{rec}}^{H} = \frac{1}{M}\sum_{j=1}^{M}\left[-\log p_G(H_j|E(H_j)) + \frac{1}{M}\left[-\log(\alpha^{(j)}) - \sum_{\substack{k\neq j \\ k=1}}^{M}\log(1-\alpha^{(k)})\right]\right]$$

reconstruction constraint

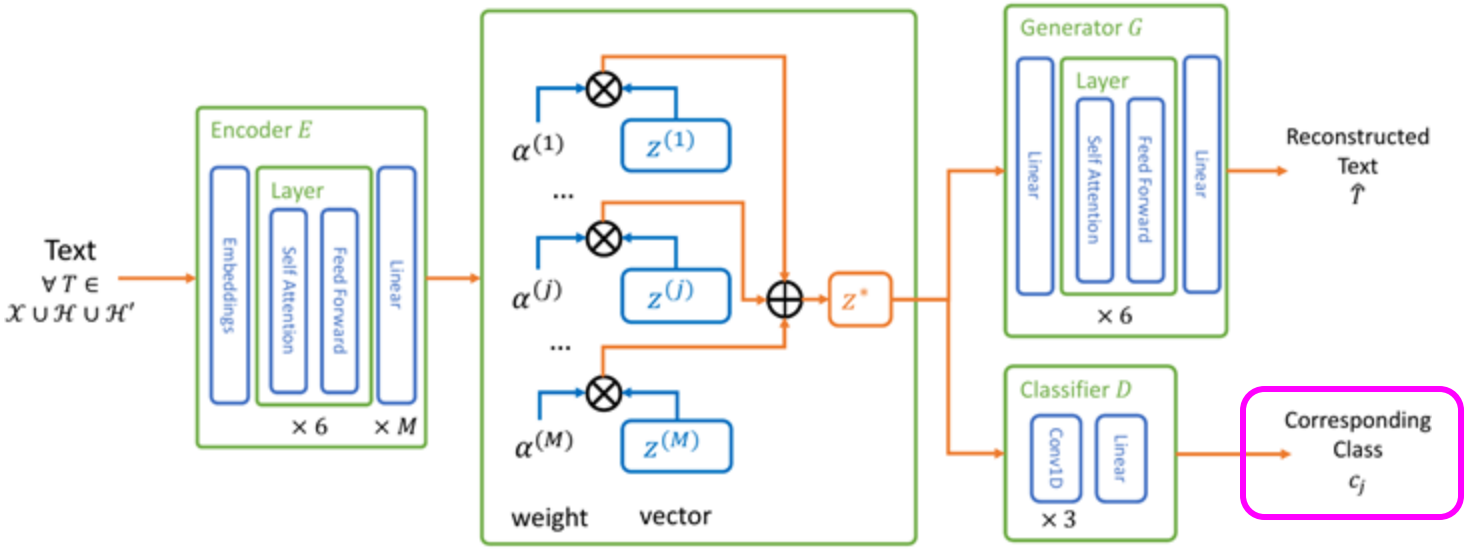
An Auto-Encoder Model



- Loss 3: text reconstruction of the phenotype subclasses' description.

$$\mathcal{L}_{rec}^{H'} = \frac{1}{|\mathcal{H}'|} \sum_{H' \in \mathcal{H}'} \left[\underbrace{-\log p_G(H'|E(H'))}_{\text{reconstruction}} + \frac{1}{M} \underbrace{\left[-\sum_{j=1}^{M} \sum_{H' \in S(H_j)} \log(\alpha^{(j)}) - \sum_{j=1}^{M} \sum_{H' \notin S(H_j)} \log(1-\alpha^{(j)}) \right]}_{\text{constraint}} \right]$$

An Auto-Encoder Model



- Loss 4: the latent vectors sampled from different priors can be classified to different classes, then the priors are thought to be 'distinct' enough.

$$\mathcal{L}_{\text{pr}} = \frac{1}{\#_T} \sum_T \sum_{j=1}^{M} \left[-\log p_D(c_j | z^{(j)} \in E(T)) \right]$$

Algorithm

Algorithm 1: The training algorithm.

Input: EHRs $\mathcal{X}$ (training set), general phenotypic abnormalities $\mathcal{H}$, and additional subclasses $\mathcal{H}'$.

Initializing θ_E, θ_G, θ_D ;

repeat

  Sample a mini-batch of B textual examples $\{T_{(i)}\}_{i=1}^{B} \subseteq \mathcal{X} \cup \mathcal{H} \cup \mathcal{H}'$;

  Get $z^*_{(i)}$ and $\{z^{(j)}_{(i)}\}_{j=1}^{M}$ by $E(T_{(i)})$;

  Reconstruct $\hat{T}_{(i)}$ by $G(z^*_{(i)})$;

  Calculate $\mathcal{L}^X_{\text{rec}}$, $\mathcal{L}^H_{\text{rec}}$, $\mathcal{L}^{H'}_{\text{rec}}$ respectively ;

  Classify $z^{(j)}_{(i)}$ by $D(z^{(j)}_{(i)})$;

  Calculate $\mathcal{L}_{\text{pr}}$ by Equation 7 ;

  Update θ_E, θ_G, θ_D by gradient descent on:

$$\mathcal{L} = \lambda_1 \mathcal{L}^X_{\text{rec}} + \lambda_2 \mathcal{L}^H_{\text{rec}} + \lambda_3 \mathcal{L}^{H'}_{\text{rec}} + \lambda_4 \mathcal{L}_{\text{pr}} \qquad (6)$$

until *convergence*;

Output: The encoder E.

Experiments - Datasets

- Discharge summaries from MIMIC-III as EHR.
  - #EHR: 52,722
  - 70% as training, 30% as testing (random split)
- Human Phenotype Ontology (HPO).
  - General phenotype $M = |\mathcal{H}| = 24$
  - Phenotype subclasses $|\mathcal{H}'| = 13795$

Experiments - Time Efficiency

A COMPARISON OF DIFFERENT METHODS. THE #RECORDS REFERS TO THE NUMBER OF TEXTUAL RECORDS USED IN THE ORIGINAL WORKS. THE TIME WAS MEASURED BY THE DURATION OF ANNOTATING 52,722 EHRS IN INFERENCE STAGE WITH A SINGLE THREAD INTEL I7-6850K 3.60GHZ AND A SINGLE NVIDIA TITAN X.

Method	Available (A) Open source (O)	#Records	Time to annotate 52,722 EHRs
OBO	A, Not O	515	1.0 hour
NCBO	A, Not O	/	36.7 hours
MetaMap	A, O	/	$\sim$ 22 days
Bio-LarK	Not A, Not O	228	/
CNN [16]	Not A, Not O	1,610	/
Ours	A, O	**52,722**	**40.2 min**

Experiments - Accuracy

THE PERFORMANCE OF ANNOTATION RESULTS COMPARED WITH THE SILVER STANDARD. ALL THE NUMBERS ARE AVERAGED ACROSS EHRS IN THE TESTING SET.

Method	Precision	Recall	F1
Random	0.5541	0.5401	0.5108
Keyword	0.6732	0.4982	0.5194
OBO	0.6817	0.5917	0.5775
NCBO	0.6782	0.5724	0.5659
MetaMap	**0.7425**	0.5231	0.5576
Ours	0.7113	**0.6805**	**0.6383**

- Silver standard

[1] https://hpo.jax.org/app/download/annotation
[2] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L.Barab´asi, "The human disease network,"Proceedings of the NationalAcademy of Sciences, vol. 104, no. 21, pp. 8685–8690, 2007.
[3] J. Park, D.-S. Lee, N. A. Christakis, and A.-L. Barab´asi, "The impact ofcellular networks on disease comorbidity,"Molecular systems biology,vol. 5, no. 1, p. 262, 2009.

Conclusion & Future Works

- A novel unsupervised deep learning framework to annotate phenotype from EHRs.
- The experiments have shown the effectiveness and efficiency of our method.
- We believe our method can provide a better indication for disease diagnosis.

- Integrate external biomedical literature from PubMed, Elsevier, etc.
- Extend to annotate all the 13k specific phenotype in HPO.
- Improve embeddings of HPO by taking both semantics and hierarchy into consideration.
- Apply to general domains.

Imperial College London

Thanks! QA?

Jingqing Zhang
Data Science Institute
Imperial College London
jz9215@imperial.ac.uk