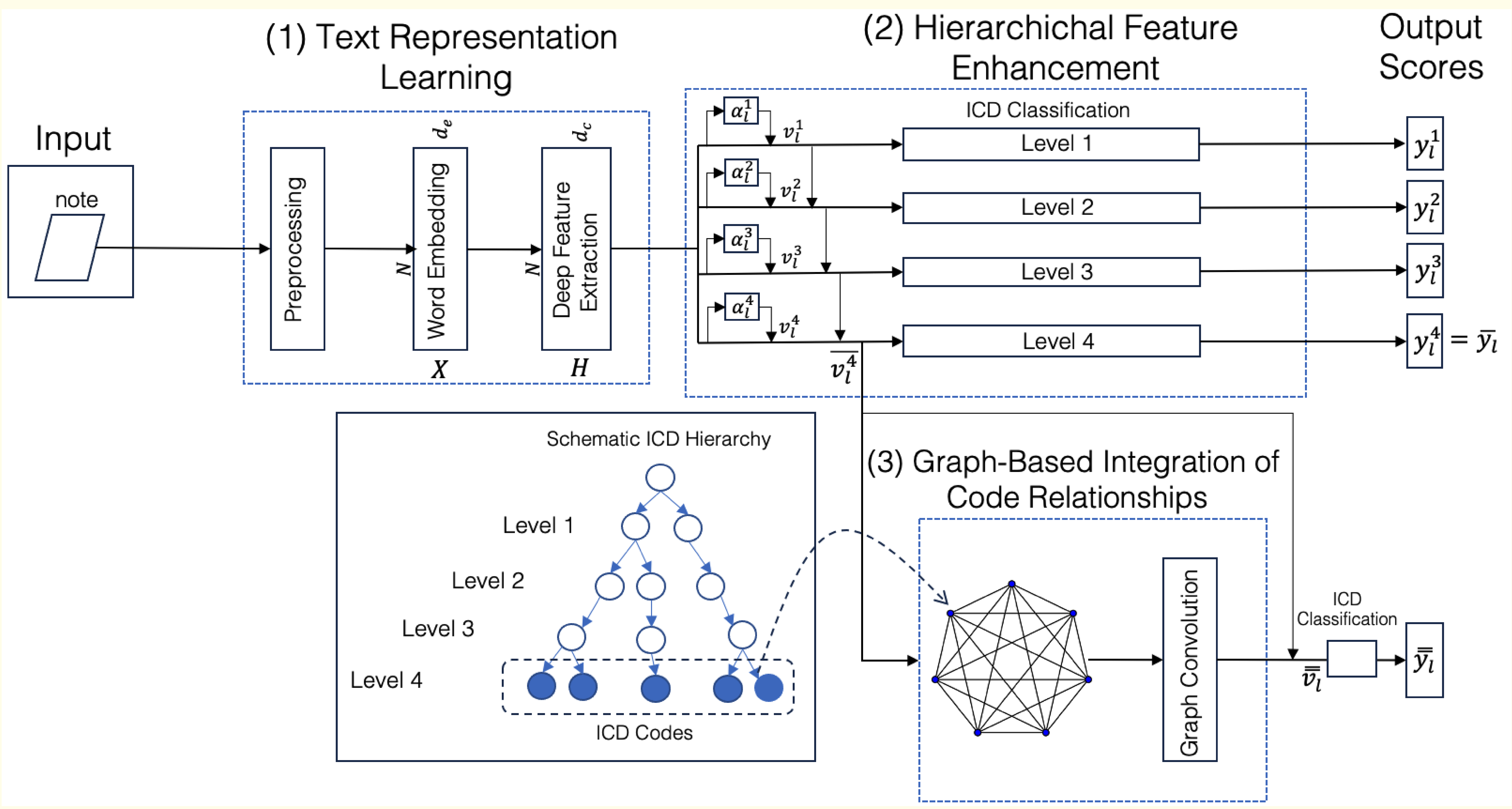
CO-OCCURRENCE GRAPH-ENHANCED HIERARCHICAL PREDICTION OF ICD CODES

Soha S. Mahdi[1], Eirini Papagiannopoulou[1], Nikos Deligiannis[1,2], Hichem Sahli[1,2]
[1] *ETRO Department, Vrije Universiteit Brussel (VUB), Brussels, Belgium*
[2] *imec, Leuven, Belgium*

Introduction

This study presents a modular approach, sequentially combining graph-based integration of ICD code co-occurrence with a hard-coded hierarchical enriched text representation drawn from the ICD ontology.

Our proposed ICD classification model architecture



Results

Mean±standard deviation derived from five different runs on the MIMIC-III dataset, compared based on f1-micro, f1-macro, and precision@8 metrics. Baseline, models with varied levels of hierarchical enhancement (HE), models with hierarchical enhancement and graph-based enhancement (HE + GBE), and a model with graph-based enhancement only are tested.

Model	Model Details	f1-micro	f1-macro	prec@8
Baseline	Model1: CAML[4]	0.5105 ± 5e-4	0.0703 ± 8e-4	0.6396 ± 7e-4
Models with HE	Model2: 2 level HE	0.5177 ± 12e-4	0.0721 ± 26e-4	0.6524 ± 10e-4
	Model3: 3 level HE	0.5221 ± 13e-4	0.0728 ± 17e-4	0.6557 ± 1e-4
	Model4: 4 level HE	0.5195 ± 11e-4	0.0728 ± 29e-4	0.6526 ± 14e-4
Models with HE and GBE	Model5: Model 4 + GBE	**0.5237 ± 8e-4**	0.07349 ± 9e-4	**0.6568 ± 13e-4**
	Model6: Model 3 + GBE	0.5195 ± 16e-4	**0.0748 ± 5e-4**	0.6519 ± 20e-4
Models with GBE	Model7: Model 1 + GBE	0.5129 ± 3e-4	0.0692 ± 17e-4	0.6401 ± 19e-4

Model1: CAML baseline (utilizing per-label attention for feature extraction). **Model2** through **Model4** integrate the hierarchical feature enhancement module, each successively integrating an additional hierarchical level compared to the Model1. **Model4** encompasses the entirety of the hierarchical levels. **Model5** integrates the graph-based code relationship module following the hierarchical enhancements introduced in Model4. Building upon the insight that Model3 outperformed Model4, we combine Model3 with the graph-based module to create **Model6**. Our analysis is completed by including **Model7**, which attaches the graph-based module directly after the baseline Model1.

Conclusion

- Our study demonstrated enhanced performance using sequentially combined modules for text feature extraction and ICD coding, outperforming CAML;
- Its modular design allows seamless integration into existing models;
- Further research could extend this approach to larger datasets and explore ICD-10 or ICD-11 applicability.

Acknowledgements

This work was partially funded by INNOVIRIS (Brussels Capital Region, Belgium) under the project: "REFINED".

References

[1] J. Mullenbach *et al.*, "Explainable Prediction of Medical Codes from Clinical Text," in *Proceedings of NAACL 2018: H L T, Vol. 1.* ACL, 2018, pp. 1101–1111.

Method

1. Text Representation Learning Each note $X \in \mathbb{R}^{N \times d_e}$, where N the number of words and d_e the dimension of the word embeddings, goes through a convolutional layer with filter size k, producing a base representation

$$\boldsymbol{H} = \widetilde{t}(\boldsymbol{W} * \boldsymbol{X} + \boldsymbol{b}), \qquad (1)$$

where $\widetilde{t}$, $W \in \mathbb{R}^{d_e \times d_r}$, and $b \in \mathbb{R}^{d_e}$ are the tanh, convolutional filters and biases.

2. Hierarchical Feature Enhancement An attention mechanism is developed for each of the four hierarchical levels. For level i, the CAML [1] attention mechanism enhances features to $v_\ell^i \in \mathbb{R}^{d_r \times 1}$, which are then combined with features from previous levels. For level $i > 1$, the output of the hierarchical module with i levels is:

$$\bar{\boldsymbol{v}_l}^{\boldsymbol{i}} \leftarrow (\boldsymbol{v}_{\boldsymbol{\ell}}^{\boldsymbol{j}} ||_{\boldsymbol{j=1}}^{\boldsymbol{i}}). \qquad (2)$$

3. Graph-based Integration of Code Relationships The graph convolution function g takes the text representation H^0 as input and outputs:

$$\boldsymbol{H}^{\boldsymbol{1}} = \boldsymbol{g}(\boldsymbol{H}^{\boldsymbol{0}}) = \boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{H}^{\boldsymbol{0}}\boldsymbol{W}^{\boldsymbol{0}}), \qquad (3)$$

where $H^0 = \bar{v_l}^4$ is the input feature representation, σ is the LeakyReLU activation, $A \in \mathbb{R}^{l \times l}$ contains edge frequency weights in the graph with l nodes (classes), and $W^0 \in \mathbb{R}^{d_r \times d_r}$ are the learnable weights. The enhanced representation for code ℓ is:

$$\bar{\bar{\boldsymbol{v}}}_{\boldsymbol{\ell}} \leftarrow (\bar{\boldsymbol{v}_{\boldsymbol{\ell}}}^{\boldsymbol{4}} || \boldsymbol{H}^{\boldsymbol{1}}). \qquad (4)$$

Classification and Loss Function The classification for each hierarchical level is: $\boldsymbol{y}_{\boldsymbol{\ell}}^{\boldsymbol{i}} = \boldsymbol{\sigma}(\boldsymbol{\beta}^{\boldsymbol{i}\top}_{\boldsymbol{\ell}} \bar{\boldsymbol{v}_{\boldsymbol{\ell}}}^{\boldsymbol{i}} + \boldsymbol{b}_{\boldsymbol{\ell}}^{\boldsymbol{i}})$, where $\bar{v_\ell}^i$ is the document representation vector for label ℓ in level i. The prediction weights $\beta_\ell^i \in \mathbb{R}^{d_r \times i}$ and the scalar offset b_ℓ^i are learned parameters specific to each label ℓ in level i. For the last classifier, $\bar{\bar{y}}_\ell = \sigma(\bar{\beta}_\ell^\top \bar{\bar{v}}_\ell + \bar{b}_\ell)$, where $\bar{\beta}_\ell \in \mathbb{R}^{d_r \times 5}$. For the multi-label classification task of hierarchical levels y_l^i and the graph-based module output $\bar{\bar{y}}_l$, the binary cross-entropy with logits loss function is used. The final loss function comprises two terms:

$$\ell = \sum_{i=1}^{4} 10^{i-3} \ell_{BCE}(\boldsymbol{y}_{\boldsymbol{\ell}}^{\boldsymbol{i}}, \boldsymbol{t}_{\boldsymbol{\ell}}^{\boldsymbol{i}}) + \ell_{BCE}(\bar{\bar{\boldsymbol{y}}}_{\boldsymbol{\ell}}, \bar{\bar{\boldsymbol{t}}}_{\boldsymbol{\ell}}),$$

where

$$\ell_{BCE}(y_\ell, t_\ell) = -t_\ell \log(y_\ell) - (1 - t_\ell)\log(1 - y_\ell).$$