Multi-label Few/Zero-shot Learning with Knowledge Aggregated from Multiple Label Graphs

Jueqing Lu[1], **Lan Du**[1*], **Ming Liu**[2], **and Joanna Dipnall**[3]
[1]Faculty of Information Technology, Monash University, Australia, VIC 3800
[2]School of Information Technology, Deakin University, Australia, VIC 3217
[3]School of Public Health and Preventive Medicine, Monash University, Australia, VIC 3800
jluu0014@student.monash.edu
{lan.du, jo.dipnall}@monash.edu
m.liu@deakin.edu.au

Abstract

Few/Zero-shot learning is a big challenge of many classifications tasks, where a classifier is required to recognise instances of classes that have very few or even no training samples. It becomes more difficult in multi-label classification, where each instance is labelled with more than one class. In this paper, we present a simple multi-graph aggregation model that fuses knowledge from multiple label graphs encoding different semantic label relationships in order to study how the aggregated knowledge can benefit multi-label zero/few-shot document classification. The model utilises three kinds of semantic information, i.e., the pre-trained word embeddings, label description, and pre-defined label relations. Experimental results derived on two large clinical datasets (i.e., MIMIC-II and MIMIC-III) and the EU legislation dataset show that methods equipped with the multi-graph knowledge aggregation achieve significant performance improvement across almost all the measures on few/zero-shot labels.

1 Introduction

Multi-label learning is a fundamental and practical problem in computer vision and natural language processing. Many tasks, such as automated medical coding (Yan et al., 2010; Rios and Kavuluru, 2018; Du et al., 2019), recommender systems (Halder et al., 2018), image classification (Chen et al., 2019; Wang et al., 2020), law study (Parikh et al., 2019; Chalkidis et al., 2019), and stance detection (Ferreira and Vlachos, 2019) can be formulated as a multi-label learning problem. Different from multi-class classification, an instance in multi-label learning is often associated with more than one class label, which makes the task even more challenging due to the combinatorial nature of the label space. i.e., the number of possible label combinations is exponential with the total number of labels.

In real-world applications, there are often insufficient or even unavailable training data of ever emerging classes (Vinyals et al., 2016; Xian et al., 2019). For instance, more than half of the International Classification of Diseases (ICD) codes are not associated with a discharge summary in the MIMIC-III dataset (Johnson et al., 2016; Rios and Kavuluru, 2018). As a solution, zero-shot learning (Xian et al., 2019; Wang et al., 2019) aims to generalize classifiers to unseen classes by leveraging various label semantics. Those classifiers are required to recognise instances of classes that have never been seen in the training set, which becomes more difficult in multi-label learning.

Moreover, the number of classes can reach hundreds of thousands. The ICD-9-CM taxonomy contains 17K diagnosis/procedure codes[1], where the majority occurs less than 10 times in MIMIC-III; the EU legislation corpus (EURLEX57X) (Chalkidis et al., 2019) contains about 7K labels, 70%of which have been assigned to less than 10 documents. The power-law distribution of labels (Liu et al., 2017; Xie et al., 2019; Song et al., 2019) leads to the few-shot learning challenge, where each label has a few training instances.

Classes come naturally with structures, which capture different relationships between individual classes. For example, codes in the ICD-9-CM taxonomy are organised in a rooted tree with edges representing is-a relationships between parents and children (Perotte et al., 2014). We can compute a code similarity graph using the code description and a code co-occurrence graph using the annotated discharge summaries in MIMIC-II/III. These two graphs can capture label relationships that are missing in the taxonomy. For example, the sim-

*Corresponding author

[1]https://www.cdc.gov/nchs/icd/icd9cm.htm

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2935–2943,
November 16–20, 2020.

ilarity graph can reveal the relationship between "hypertensive chronic kidney disease" and "acute kidney failure"; the co-occurrence graph can give us information about that "coronary atherosclerosis of native coronary artery" frequently co-occurs with "coronary arteriography using two catheters". It has been shown that ignoring this structured information and assuming all classes to be mutually exclusive are insufficient (Zhao et al., 2018; Gaure and Rai, 2017; Kavuluru et al., 2015).

In this paper, we present a simple but effective multi-graph knowledge aggregation model that can transform and fuse the structural information from multiple label graphs while utilising three kinds of semantics: the pre-trained word embeddings, label description, and the label relations. To demonstrate its efficacy, we adapt the model as a sub-module to several existing neural architectures (Rios and Kavuluru, 2018; Chalkidis et al., 2019) for multi-label few/zero-shot learning. However, this model can work as a self-contained module and be flexibly adapted to most existing multi-label learning models (Xie et al., 2019; Li and Yu, 2020) that use GCNs to leverage the label structures. Experiments on three real-world datasets show that neural classifiers equipped with our multi-graph knowledge aggregation model can significantly improve the few/zero-shot classification performance.

2 Related Work

Leveraging structural label information via GCNs (Kipf and Welling, 2017) has become a promising approach of tackling the few/zero-shot problem, attracting increasing attention in recent years. Wang et al. (2018); Kampffmeyer et al. (2019), and Chen et al. (2017) have used GCNs to learn visual classifiers for multi-class image classification. These ideas can be generalised to multi-label learning (Lee et al., 2018; Chen et al., 2019; Do et al., 2019; Wang et al., 2020; You et al., 2020). However, none of these methods can be directly adapted to multi-label few/zero-shot text classification. Using the label-wise attention mechanism (Mullenbach et al., 2018; Xiao et al., 2019), Rios and Kavuluru (2018) introduced an attention-based CNN to convert each document into a feature matrix, each row of which is a label-specific document feature vector. The multi-label document classifiers were learned from a GCN over the label hierarchy. While considering only the efficiency of the document encoder, Chalkidis et al. (2019); Li and Yu (2020); Xie et al. (2019) further proposed to replace the simple CNN with BIGRU, multi-filter residual CNN and densely-connected CNN respectively. In contrast, our work focuses on the learning of the classifiers from multiple label graphs. Existing work on multiple graphs learning often proposed to either fuse multiple graphs before fed into a GCN (Khan and Blumenstock, 2019; Wang et al., 2020) or consider the multi-dimensionality of graphs (Ma et al., 2019; Wu et al., 2019) for only note classification/link prediction.

3 Learning with Knowledge Aggregation

Problem Formulation Let $\mathcal{C}_S$ and $\mathcal{C}_U$ be disjoint sets of seen and unseen labels. $\mathcal{C}_S$ is further divided into frequent labels $\mathcal{C}_S^R$ and few-shot labels $\mathcal{C}_S^F$ such that $\mathcal{C}_S = \mathcal{C}_S^R \cup \mathcal{C}_S^F$. Given a training set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$, where $\mathbf{x}_i$ indicates the i-th document and $\mathbf{y}_i \subset \mathcal{C}_S$ is the subset of labels assigned to $\mathbf{x}_i$, the goal is to predict $\hat{\mathbf{y}}_i$ for each test document in generalised zero-shot settings (Xian et al., 2019), where $\hat{\mathbf{y}}_i$ is a subset of $\mathcal{C}_S \cup \mathcal{C}_U$. Note that: *i)* every label has a description; *ii)* the label relationships encoded in graphs can be computed from various resources; *iii)* documents associated with any label from $\mathcal{C}_U$ are excluded from training.

Document Encoder with Label-wise Attention According to the characteristic of different datasets, different document encoders ϕ can be used to generate the document representation, i.e., $\mathbf{F}_i = \phi(\mathbf{x}_i)$. For a corpus, like EURLEX57X, where the average document length is in hundreds, one can consider Bi-GRU/LSTM, HAN (Yang et al., 2016), BERT (Devlin et al., 2019), etc. For a corpus, like MIMIC-II/III, where the discharge summaries contain multiple long and heterogeneous medical narratives, the CNN-based encoders have shown prominet performance, like those discussed in Section 2. The size of $\mathbf{F}_i \in \mathbb{R}^{n \times u}$ varies, depending on the encoder. For BERT, n is the number of words and u is the size of the output layer of BERT; for CNNs, n is the number of s-grams generated by CNNs with a filter size s and u the number of filters.

In addition, we create label embeddings $\mathbf{v}_l$ by TF-IDF weighted average of pre-trained word embeddings (Chen et al., 2017) according to the label description, and use those label embeddings to compute the label-wise attention (Mullenbach et al., 2018; Rios and Kavuluru, 2018) for each

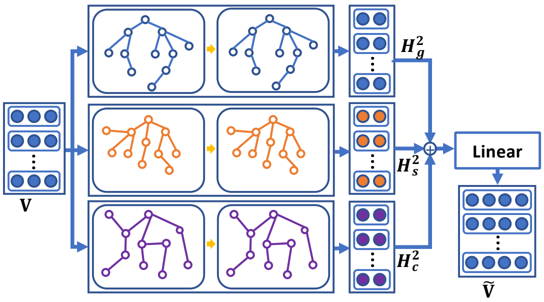

Figure 1: Multi-graph knowledge aggregation

document $\mathbf{x}_i$ as follows:

$$\mathbf{a}_{i,l} = \text{softmax}(\tanh(\mathbf{F}_i \mathbf{W}_0 + \mathbf{b}_0)\mathbf{v}_l) \quad (1)$$
$$\mathbf{z}_{i,l} = \mathbf{a}_{i,l}^T \mathbf{F}_i, \quad (2)$$

where $\mathbf{W}_0 \in \mathbb{R}^{u \times d}$, $\mathbf{b}_0 \in \mathbb{R}^d$. The attention is to capture how different parts of texts are relevant to different classes.

Knowledge Aggregation from Multi-Graphs (KAMG) We consider the label hierarchy ($\mathbf{A}_g$) given by the class taxonomy, the semantic similarity graph ($\mathbf{A}_s$) computed from their descriptions, and the label co-occurrence graph ($\mathbf{A}_c$) extracted for $\mathcal{C}_S$ from the training data, although our method can be generated to more label graphs. Let $\mathbf{A} \in \mathbb{R}^{|\mathcal{C}_S| \times |\mathcal{C}_S|}$ be any of the three label graphs, $\mathbf{V} \in \mathbb{R}^{L \times d}$ be the label embedding matrix, a two-layer GCN is applied to each graph as follows:

$$\mathbf{H}^1 = \sigma(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{V}\mathbf{W}_1) \quad (3)$$
$$\mathbf{H}^2 = \sigma(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{H}^1\mathbf{W}_2) \quad (4)$$

where $\mathbf{D}_{i,i} = \sum_j A_{i,j}$ is a degree matrix of $\mathbf{A}$, $\mathbf{W}^1 \in \mathbb{R}^{d \times q}$ and $\mathbf{W}^2 \in \mathbb{R}^{q \times p}$ are two weight matrices, $\mathbf{H}^1$ and $\mathbf{H}^2$ indicate the hidden states and outputs respectively, σ is the non-linear activation function, a rectified linear unit (ReLU) in our case.

Different from Rios and Kavuluru (2018); Xie et al. (2019), we feed a two-layer GCN to each of the three graphs and generate three sets of label embeddings: $\mathbf{H}_g^2$, $\mathbf{H}_s^2$ and $\mathbf{H}_c^2$, which are supposed to capture different semantic relations between labels. A linear layer is then used to fuse the three types of label embeddings:

$$\tilde{\mathbf{v}}_l = f([\mathbf{h}_{g,l}^2, \mathbf{h}_{s,l}^2, \mathbf{h}_{c,l}^2], \mathbf{W}_3) \quad (5)$$

where $\mathbf{W}_3 \in \mathbb{R}^{3p \times \tilde{q}}$, and $\tilde{\mathbf{v}}_l \in \mathbb{R}^{\tilde{q}}$. We acknowledge that it is also worth trying the techniques used in multi-model learning (Kiela et al., 2018), which is subject to future work. Figure 1 visualises the multi-graph knowledge aggregation process.

We concatenate both $\mathbf{v}_l$ with $\tilde{\mathbf{v}}_l$ to form the final text classifiers as $\bar{\mathbf{v}}_l = [\mathbf{v}_l, \tilde{\mathbf{v}}_l]$, $\bar{\mathbf{v}}_l \in \mathbb{R}^{d+\tilde{q}}$. The label-wise document embeddings ($\mathbf{z}_{i,l}$) are projected onto the same space as $\bar{\mathbf{v}}_l$ via a simple non-linear transformation as

$$\bar{\mathbf{z}}_{i,l} = \text{ReLU}(\mathbf{W}_4 \mathbf{z}_{i,l} + \mathbf{b}_4) \quad (6)$$

where $\mathbf{W}_4 \in \mathbb{R}^{(d+\tilde{q}) \times u}$ and $\mathbf{b}_4 \in \mathbb{R}^{(d+\tilde{q})}$. The prediction for each label l is generated with $\hat{y}_{i,l} = \text{sigmoid}(\bar{\mathbf{z}}_{i,l}^T \bar{\mathbf{v}}_l)$. The model is optimised via a multi-label binary cross-entropy loss. Although we used three label graphs (label hierarchy, similarity and co-occurrence) to demonstrate the advantage of aggregating knowledge from multi-graphs, the model itself is general enough to be applied to other datasets where there exist multiple label graphs.

Zero-Shot Classification For zero-shot prediction, we extend $\mathbf{A} \in \mathbb{R}^{|\mathcal{C}_S| \times |\mathcal{C}_S|}$ to $\tilde{\mathbf{A}} \in \mathbb{R}^{(|\mathcal{C}_S|+|\mathcal{C}_U|) \times (|\mathcal{C}_S|+|\mathcal{C}_U|)}$, so that the new graph can encode the relationship between unseen and seen classes. All labels will be optimized simultaneously during the training stage as in (Rios and Kavuluru, 2018). Note that $\mathbf{A}_c$ counts only the co-occurrence of seen classes.

4 Experiments

In this section, several experiments were conducted to evaluate the efficacy of KAMG in classifying discharge summaries and legislative documents. We compared our methods with several state-of-the-art multi-label classifiers in a few/zero-shot setting, and studied how KAMG behaves by varying label graphs in a set of ablation experiments.

Datasets We used two benchmark medical datasets (MIMIC II and III) and the EU legislation dataset (EURLEX57K) to evaluate our method in the few/zero-shot settings. Statistics of these datasets are shown in Table 1. Following Rios and Kavuluru (2018); Chalkidis et al. (2019), we split the datasets in such a way that 1) zero-shot labels (i.e., unseen) do not have any instances in training; 2) few-shot labels (i.e., less frequent labels) were defined as those whose frequencies in the training set are less than or equal to 5 for MIMIC-II and MIMI-III and 50 for EURLEX57K. The 200-dimensional word embeddings pre-trained on PubMed and MIMIC-III (Zhang et al., 2019; Chen et al., 2019) were used for MIMIC-II/III, and 200-dimensional word embeddings pre-trained on law corpra provided by Chalkidis et al. (2019) were used for EURLEX57k.

Experiment settings and metrics For MIMIC-II/III, we used the NeuralClassifier (Liu et al., 2019) as a base framework to implement our methods. We used 200 filters with kernel size 10 to setup

	Docs						# Labels		
Dataset	#Train	#Dev	#Test	Avg # tokens	Avg # labels	Voc Size	Frequent	Few	Zero
MIMIC-II	17,593	1,955	2,200	1,350	9	55,237	1,844	2,745	361
MIMIC-III	47,718	1,631	3,372	1,931	15	104,656	4,204	4,115	203
EURLEX57K	45,000	6,000	6,000	727	5	169,439	746	3,362	163

Table 1: Dataset statistics

		Frequent		Few		Zero		Overall	
		R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
MIMIC-II	CNN (Kim, 2014)	0.346	0.465	0.032	0.018	-	-	0.335	0.460
	RCNN (Lai et al., 2015)	0.386	0.505	0.081	0.047	-	-	0.373	0.498
	CAML (Mullenbach et al., 2018)	0.386	0.508	0.078	0.043	0.021	0.012	0.371	0.501
	DR-CAML (Mullenbach et al., 2018)	0.383	0.502	0.075	0.044	0.028	0.016	0.368	0.495
	ZACNN (Rios and Kavuluru, 2018)	0.445	0.562	0.180	0.114	0.362	0.225	0.424	0.551
	ZAGCNN (Rios and Kavuluru, 2018)	**0.471**	**0.591**	0.219	0.139	0.382	0.231	**0.452**	**0.583**
	ACNN-KAMG	**0.471**	**0.591**	**0.259**	**0.166**	**0.462**	**0.296**	0.451	0.582
MIMIC-III	CNN (Kim, 2014)	0.366	0.632	0.074	0.044	-	-	0.361	0.631
	RCNN (Lai et al., 2015)	0.376	0.648	0.118	0.070	-	-	0.370	0.646
	CAML (Mullenbach et al., 2018)	0.422	0.711	0.104	0.073	0.067	0.029	0.415	0.709
	DR-CAML (Mullenbach et al., 2018)	0.416	0.699	0.105	0.064	0.038	0.018	0.409	0.697
	ZACNN (Rios and Kavuluru, 2018)	0.405	0.684	0.207	0.104	0.457	0.222	0.372	0.654
	ZAGCNN (Rios and Kavuluru, 2018)	0.427	0.713	0.258	0.130	0.512	0.253	0.394	0.685
	ACNN-KAMG	**0.434**	**0.724**	**0.295**	**0.195**	**0.553**	**0.358**	**0.427**	**0.722**

Table 2: Multi-label classification results on MIMIC-II and MIMIC-III. Bold figures indicate the best results for each score.

	Frequent		Few		Zero		Overall	
	R@5	nDCG@5	R@5	nDCG@5	R@5	nDCG@5	R@5	nDCG@5
BIGRU-LWAN (Chalkidis et al., 2019)	*0.755*	*0.819*	*0.661*	*0.618*	0.029	0.019	*0.692*	*0.796*
ZERO-CNN-LWAN (Chalkidis et al., 2019)	0.683	0.745	0.494	0.454	0.321	0.264	0.617	0.717
ZERO-BIGRU-LWAN (Chalkidis et al., 2019)	0.716	0.780	0.560	0.510	0.438	0.345	0.648	0.752
AGRU-KAMG	**0.731**	**0.795**	**0.563**	**0.518**	**0.528**	**0.414**	**0.661**	**0.766**

Table 3: Multi-label classification results on EURLEX57K. Bold figures indicate the best results for each score among the three models designed specifically for zero-shot learning. Italics indicate the best results overall.

the CNNs by following Rios and Kavuluru (2018) and the GCNs' hidden layer size was set to 200. For EURLEX57K, we leveraged Chalkidis et al. (2019)'s code, and used the one-layer BiGRU with hidden dimension 100 as reported in their paper. The size of the GCNs' hidden states was set to 200. Moreover, the dropout rate was set to 0.2, 0.1 for MIMIC-II/III and EURLEX57K respectively and applied after the embedding layer. Adam optimizer (i.e., learning rate: 0.001 for CNN and 0.0003 for BIGRU) was used to train all the models. All experiments were run with one NVIDIA GPU V100.

We report a variety of ranking metrics, including Recall@K and nDCG@K. We argue that the ranking metrics are more preferable for few/zero-shot label without introducing significant bias towards frequent labels; they are more inline with the human annotation process, like the ICD coding, where clinicians often review a limited number of candidate codes. K was set to 10 for MiMIC-II/III and 5 for URLEX57K.

Results on MIMIC-II/III We compared KAMG, which uses all three label graphs ($\mathbf{H}_g$, $\mathbf{H}_s$ and $\mathbf{H}_c$), with the following baselines: CNN, RCNN (the best model in Liu et al. (2019)), CAML, DR-CAML, ZACNN and ZAGCNN. Table-2 shows the performance of all those models. KAMG outperforms the other models in all the metrics across almost all the settings on both datasets with a notable margin, due to our multi-graph knowledge aggregation model. Specifically, while classifying zero-shot labels, ACNN-KAMG outperforms ZAGCNN, which uses only the label hierarchy (i.e., $\mathbf{H}_g$), by 8% in R@10 and 6.5% in nDCG@10 on MIMIC-II and 4.1% in R@10 and 10.5% in nDCG@10 on MIMIC-III. Similarly, ACNN-KAMG gains 4% in R@10 and 2.7% in nDCG@10 on MIMIC-II and 3.7% in R@10 and 6.5% in nDCG@10 on MIMIC-III over ZAGCNN on few-shot labels.

Results on EURLEX57K We further compared AGRU-KAMG with with BIGRU-LAWN, ZERO-CNN-LAWN, and ZERO-BIGRU-LAWN, which are the best performing models using label-wise attention on few/zero-shot labels in (Chalkidis et al., 2019). We implemented AGRU-KAMG by directly modifying ZERO-BIGRU-LAWN's published code. Results in Table 3 show AGRU-KAMG performs significantly better than ZERO-BIGRU-LAWN on zero-shot labels by gaining 9.0% improvement in R@5 and 6.9% in nDCG@5, and comparably with ZERO-BIGRU-LAWN on few-shot labels. BIGRU-LAWN exhibits strong performance on frequent/few-shot labels, which

		Frequent		Few		Zero		Overall	
		R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
MIMIC-II	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	**0.477**	**0.597**	**0.274**	**0.180**	**0.451**	**0.301**	**0.457**	**0.588**
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.470	0.587	0.235	0.151	0.418	0.273	0.450	0.578
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	**0.476**	**0.596**	**0.277**	**0.177**	**0.454**	**0.282**	**0.456**	**0.586**
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.467	0.586	0.236	0.152	0.417	0.267	0.448	0.577
MIMIC-III	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	**0.435**	**0.725**	**0.293**	**0.193**	0.530	**0.346**	**0.428**	**0.723**
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.426	0.712	0.256	0.130	**0.540**	0.273	0.393	0.684
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	**0.432**	**0.721**	**0.284**	**0.192**	**0.560**	**0.370**	**0.425**	**0.720**
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.422	0.707	0.245	0.123	0.521	0.265	0.392	0.680

Table 4: The comparison of the knowledge fusion before and after GCN on MIMIC-II and MIMIC-III. Bold figures indicate the best results for each score

	MIMIC-II				MIMIC-III			
	Few		Zero		Few		Zero	
	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
ACNN-KAMG ($\mathbf{H}_g$)	0.219	0.139	0.382	0.231	0.258	0.130	0.512	0.253
ACNN-KAMG ($\mathbf{H}_s$)	0.245	0.157	0.437	0.272	0.258	0.130	0.524	0.258
ACNN-KAMG ($\mathbf{H}_c$)	0.248	0.157	0.424	0.267	0.252	0.130	0.518	0.256
ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.257	0.161	0.439	0.286	0.252	0.138	0.533	0.267
ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.274	**0.180**	0.451	**0.301**	0.293	0.193	0.530	0.346
ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	**0.277**	0.177	0.454	0.282	0.284	0.192	**0.560**	**0.370**
ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.259	0.166	**0.462**	0.296	**0.295**	**0.195**	0.553	0.358

Table 5: Ablation study on MIMIC-II and MIMIC-III. We ran ACNN-KAMG with different combinations of the three graphs in the few/zero-shot setting. Bold figures indicate the best results for each score.

	Few		Zero	
	R@5	nDCG@5	R@5	nDCG@5
AGRU-KAMG ($\mathbf{H}_g$)	0.474	0.431	0.472	0.363
AGRU-KAMG ($\mathbf{H}_s$)	0.508	0.464	0.484	0.382
AGRU-KAMG ($\mathbf{H}_c$)	0.503	0.459	0.491	0.381
AGRU-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.554	0.509	0.499	0.397
AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.550	0.504	0.480	0.381
AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.554	0.507	0.517	**0.422**
AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	**0.563**	**0.518**	**0.528**	0.414

Table 6: Ablation study on EURLEX57K. We ran AGRU-KAMG with different combinations of the three graphs in the few/zero-shot setting. Bold figures indicate the best results for each score.

is inline with Chalkidis et al. (2019)'s finding. This could be attributed to the fine-tuning of label embeddings in the learning process. In contrast, AGRU-KAMG has label embeddings fixed to those computed from pretrained embedding in order to leverage label description in the zero-shot setting.

Results on pre/post-GCN fusion Table 4 shows the performance difference between the following two graph fusion methods: 1) merging two label graphs into one graph, and then feeding it into one GCN (Ma et al., 2019; Wang et al., 2020), and 2) our method, where two graphs were fed into two GCNs and then fused together. The results showed that our method performs much better than the pre-GCN fusion method.

Results on using different combinations of label graphs We further conducted a set of ablation experiments based on the use of different combinations of label graphs to study how the performance of KAMG varies while using different graphs in both few and zero-shot settings. The results in Tables 5 and 6 show that i) KAMG performs better with multiple graphs than with a single graph overall, which demonstrates it is beneficial to aggregate information from multiple graphs; ii) graphs contribute differently to the classification performance, the ICD taxonomy plays an important role while being used in conjunction with the other graphs, and the three graphs work complementary to each other on EURLEX57K.

5 Conclusion

We have proposed a multi-graph aggregation method that can effectively fuse knowledge from multiple label graphs. Experiments on MIMIC-II/III and EURLEX57K have shown that the classifiers derived from the multi-graph aggregation have achieved substantial performance improvements particularly on few/zero-shot labels. As future work, we will further study our method's ability of extreme multi-label learning (Bhatia et al., 2016) and different document encoders.

Acknowledgments

We thank anonymous reviewers for their valuable comments.

References

K. Bhatia, K. Dahiya, H. Jain, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322.

Meihao Chen, Zhuoru Lin, and Kyunghyun Cho. 2017. Graph convolutional networks for classification with a structured label space. *arXiv preprint arXiv:1710.04908*.

Q. Chen, Y. Peng, and Z. Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Kien Do, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2019. Attentional multilabel learning over graphs: a message passing approach. *Machine Learning*, 108(10):1757–1781.

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11):1279–1285.

William Ferreira and Andreas Vlachos. 2019. Incorporating label dependencies in multilabel stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6350–6354.

Abhilash Gaure and Piyush Rai. 2017. A probabilistic framework for zero-shot multi-label learning. In *UAI*.

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2018. Cold start thread recommendation as extreme multi-label classification. In *Proceedings of the The Web Conference*, pages 1911–1918.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. 2019. Rethinking knowledge graph propagation for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 11487–11496.

Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65(2):155–166.

Muhammad Raza Khan and Joshua E Blumenstock. 2019. Multi-GCN: Graph convolutional networks for multi-view networks, with applications to global poverty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 606–613.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient large-scale multi-modal classification. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5198–5204.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2267–2273.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the Thirty-forth AAAI Conference on Artificial Intelligence*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124.

Liqun Liu, Funan Mu, Pengyu Li, Xin Mu, Jing Tang, Xingsheng Ai, Ran Fu, Lifeng Wang, and Xing Zhou. 2019. NeuralClassifier: An open-source neural hierarchical multi-label text classification toolkit. In *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics: System Demonstrations.

Yao Ma, Suhang Wang, Chara C Aggarwal, Dawei Yin, and Jiliang Tang. 2019. Multi-dimensional graph convolutional networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 657–665.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1101–1111.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1642–1652.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric Xing. 2019. Generalized zero-shot ICD coding. *arXiv preprint arXiv:1909.13154*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3637–3645.

Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6857–6866.

Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-label classification with label graph superimposing. In *Proceedings of the Thirty-forth AAAI Conference on Artificial Intelligence*.

Man Wu, Shirui Pan, Lan Du, Ivor Tsang, Xingquan Zhu, and Bo Du. 2019. Long-short distance aggregation networks for positive unlabeled graph learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2157–2160.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41:2251–2265.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 466–475.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 649–658.

Yan Yan, Glenn Fung, Jennifer G. Dy, and Romer Rosales. 2010. Medical coding classification by leveraging inter-code relationships. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 193–202.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. 2020. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the Thirty-forth AAAI Conference on Artificial Intelligence*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.

He Zhao, Piyush Rai, Lan Du, and Wray Buntine. 2018. Bayesian multi-label learning with sparse features and labels, and label co-occurrences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1943–1951.

Appendices

Tables 7, 9, 10 and 8 present a full set of experiments results computed with different metrics, including, Recall@K, Precision@K, Recall-Precision@K, nDCG@K. All the experiments were run on one NVIDIA GPU V100.

		Frequent			Few			Zero			Overall		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MIMIC-II	CNN	0.080	0.253	0.346	0.005	0.021	0.032	-	-	-	0.077	0.245	0.335
	RCNN	0.086	0.277	0.386	0.015	0.048	0.081	-	-	-	0.083	0.267	0.372
	CAML	0.082	0.278	0.386	0.014	0.043	0.078	0.004	0.014	0.021	0.079	0.267	0.371
	DR-CAML	0.080	0.276	0.383	0.016	0.046	0.075	0.005	0.019	0.028	0.077	0.265	0.368
	ZACNN	0.086	0.308	0.445	0.050	0.126	0.180	0.101	0.262	0.362	0.082	0.294	0.424
	ZAGCNN	0.089	0.323	0.471	0.060	0.161	0.219	0.102	0.267	0.382	0.085	0.309	0.452
	ACNN-KAMG($\mathbf{H}_g$)	0.089	0.319	0.467	0.066	0.172	0.235	0.141	0.302	0.402	0.085	0.305	0.448
	ACNN-KAMG($\mathbf{H}_s$)	0.088	0.322	0.469	0.069	0.178	0.245	0.126	0.315	0.437	0.084	0.308	0.449
	ACNN-KAMG($\mathbf{H}_c$)	0.088	0.323	0.474	0.068	0.178	0.247	0.127	0.305	0.424	0.084	0.309	0.454
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.088	0.320	0.470	0.067	0.171	0.235	0.140	0.323	0.418	0.084	0.307	0.450
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.088	0.319	0.467	0.068	0.177	0.236	0.136	0.308	0.416	0.084	0.306	0.448
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.090	0.325	0.477	0.083	0.203	0.274	0.163	0.345	0.451	0.086	0.311	0.457
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.091	0.325	0.476	0.077	0.200	0.277	0.130	0.323	0.454	0.086	0.311	0.456
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.091	0.324	0.475	0.067	0.177	0.248	0.137	0.343	0.447	0.086	0.310	0.454
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.089	0.322	0.471	0.072	0.188	0.259	0.145	0.342	0.462	0.085	0.309	0.451
MIMIC-III	CNN	0.061	0.240	0.366	0.017	0.051	0.074	-	-	-	0.060	0.236	0.361
	RCNN	0.063	0.247	0.376	0.027	0.080	0.118	-	-	-	0.062	0.243	0.370
	CAML	0.066	0.267	0.422	0.038	0.084	0.104	0.002	0.036	0.067	0.065	0.262	0.415
	DR-CAML	0.065	0.263	0.416	0.026	0.073	0.105	0.003	0.016	0.038	0.063	0.258	0.409
	ZACNN	0.064	0.256	0.405	0.008	0.140	0.207	0.007	0.309	0.457	0.063	0.241	0.372
	ZAGCNN	0.065	0.266	0.427	0.006	0.181	0.258	0.007	0.367	0.512	0.064	0.252	0.394
	ACNN-KAMG($\mathbf{H}_s$)	0.065	0.262	0.420	0.004	0.184	0.258	0.007	0.376	0.524	0.063	0.247	0.385
	ACNN-KAMG($\mathbf{H}_c$)	0.065	0.262	0.419	0.007	0.171	0.252	0.007	0.374	0.518	0.063	0.245	0.382
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.065	0.265	0.426	0.009	0.181	0.256	0.007	0.401	0.540	0.064	0.251	0.393
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.065	0.263	0.422	0.008	0.166	0.245	0.007	0.397	0.521	0.064	0.250	0.392
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.066	0.271	0.435	0.101	0.224	0.293	0.172	0.412	0.530	0.065	0.266	0.428
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.066	0.270	0.432	0.103	0.216	0.284	0.194	0.449	0.560	0.065	0.265	0.425
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.066	0.268	0.423	0.052	0.192	0.280	0.021	0.386	0.566	0.065	0.263	0.414
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.066	0.271	0.434	0.096	0.231	0.295	0.180	0.417	0.553	0.065	0.266	0.427
EU	AGRU-KAMG ($\mathbf{H}_g$)	0.229	0.696	0.836	0.282	0.474	0.550	0.226	0.472	0.551	0.194	0.625	0.762
	AGRU-KAMG ($\mathbf{H}_c$)	0.232	0.708	0.847	0.303	0.503	0.585	0.254	0.491	0.574	0.196	0.636	0.775
	AGRU-KAMG ($\mathbf{H}_s$)	0.231	0.707	0.847	0.305	0.508	0.586	0.258	0.484	0.593	0.197	0.636	0.776
	AGRU-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.237	0.726	0.868	0.316	0.554	0.630	0.267	0.499	0.606	0.201	0.656	0.796
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.238	0.727	0.864	0.333	0.550	0.631	0.257	0.480	0.569	0.201	0.656	0.795
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.238	0.727	0.868	0.335	0.554	0.628	0.298	0.517	0.641	0.201	0.657	0.799
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.238	0.731	0.869	0.342	0.563	0.643	0.268	0.528	0.635	0.201	0.661	0.801

Table 7: Recall@k results on MIMIC-II, MIMIC-III and EURLEX57K (EU) datasets

		Frequent			Few			Zero			Overall		
		nDCG@1	nDCG@5	nDCG@10	nDCG@1	nDCG@5	nDCG@10	nDCG@1	nDCG@5	nDCG@10	nDCG@1	nDCG@5	nDCG@10
MIMIC-II	CNN	0.712	0.538	0.465	0.007	0.014	0.018	-	-	-	0.711	0.536	0.460
	RCNN	0.739	0.574	0.505	0.022	0.035	0.047	-	-	-	0.738	0.572	0.498
	CAML	0.727	0.578	0.508	0.018	0.031	0.043	0.004	0.009	0.012	0.726	0.576	0.501
	DR-CAML	0.713	0.571	0.502	0.023	0.034	0.044	0.005	0.013	0.016	0.712	0.569	0.495
	ZACNN	0.752	0.619	0.562	0.066	0.095	0.114	0.114	0.191	0.225	0.750	0.615	0.551
	ZAGCNN	0.778	0.648	0.591	0.077	0.119	0.139	0.118	0.193	0.231	0.777	0.645	0.583
	ACNN-KAMG($\mathbf{H}_g$)	0.777	0.641	0.586	0.084	0.128	0.151	0.160	0.231	0.264	0.776	0.638	0.578
	ACNN-KAMG($\mathbf{H}_s$)	0.772	0.644	0.588	0.090	0.133	0.157	0.143	0.231	0.272	0.770	0.641	0.578
	ACNN-KAMG($\mathbf{H}_c$)	0.772	0.645	0.591	0.088	0.133	0.157	0.141	0.227	0.267	0.770	0.642	0.581
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.770	0.642	0.587	0.086	0.129	0.151	0.155	0.241	0.273	0.769	0.639	0.578
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.769	0.641	0.585	0.087	0.132	0.152	0.153	0.231	0.267	0.768	0.638	0.577
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.784	0.652	0.597	0.109	0.155	0.180	0.186	0.266	0.301	0.783	0.649	0.588
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.785	0.650	0.596	0.100	0.150	0.177	0.146	0.238	0.282	0.784	0.647	0.586
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.785	0.649	0.595	0.085	0.132	0.157	0.159	0.251	0.286	0.783	0.646	0.585
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.780	0.647	0.591	0.092	0.141	0.166	0.165	0.256	0.296	0.778	0.644	0.581
MIMIC-III	CNN	0.826	0.720	0.632	0.020	0.036	0.044	-	-	-	0.826	0.719	0.631
	RCNN	0.845	0.739	0.648	0.034	0.057	0.070	-	-	-	0.845	0.738	0.646
	CAML	0.884	0.788	0.711	0.045	0.066	0.073	0.007	0.019	0.029	0.884	0.787	0.709
	DR-CAML	0.859	0.775	0.699	0.032	0.053	0.064	0.005	0.010	0.018	0.859	0.775	0.697
	ZACNN	0.858	0.762	0.684	0.010	0.081	0.104	0.007	0.173	0.222	0.858	0.748	0.654
	ZAGCNN	0.875	0.786	0.713	0.007	0.103	0.130	0.007	0.205	0.253	0.875	0.774	0.685
	ACNN-KAMG($\mathbf{H}_s$)	0.872	0.778	0.703	0.005	0.105	0.130	0.007	0.210	0.258	0.872	0.765	0.673
	ACNN-KAMG($\mathbf{H}_c$)	0.873	0.778	0.703	0.008	0.098	0.126	0.007	0.209	0.256	0.873	0.761	0.668
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.874	0.784	0.712	0.009	0.105	0.130	0.007	0.227	0.272	0.873	0.773	0.683
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.873	0.780	0.707	0.009	0.096	0.123	0.007	0.223	0.265	0.873	0.769	0.680
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.885	0.797	0.725	0.118	0.169	0.193	0.190	0.307	0.346	0.885	0.797	0.723
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.883	0.795	0.721	0.120	0.169	0.192	0.215	0.333	0.370	0.882	0.794	0.719
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.884	0.792	0.713	0.059	0.128	0.159	0.028	0.221	0.280	0.884	0.791	0.709
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.882	0.797	0.724	0.109	0.172	0.195	0.203	0.313	0.358	0.882	0.796	0.722
EU	AGRU-KAMG ($\mathbf{H}_g$)	0.857	0.760	0.805	0.415	0.431	0.460	0.247	0.363	0.388	0.862	0.729	0.760
	AGRU-KAMG ($\mathbf{H}_c$)	0.865	0.771	0.816	0.444	0.459	0.490	0.272	0.381	0.410	0.871	0.740	0.772
	AGRU-KAMG ($\mathbf{H}_s$)	0.866	0.771	0.815	0.447	0.464	0.493	0.276	0.382	0.420	0.873	0.740	0.773
	AGRU-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.881	0.790	0.834	0.496	0.509	0.538	0.285	0.397	0.432	0.889	0.761	0.793
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.882	0.791	0.834	0.489	0.504	0.534	0.267	0.381	0.409	0.888	0.761	0.793
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.884	0.792	0.837	0.491	0.507	0.535	0.323	0.422	0.462	0.891	0.763	0.796
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.883	0.795	0.839	0.504	0.518	0.548	0.290	0.414	0.447	0.891	0.766	0.798

Table 8: nDCG@k results on MIMIC-II, MIMIC-III and EURLEX57K (EU) datasets

		Frequent			Few			Zero			Overall		
		P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
MIMIC-II	CNN	0.712	0.478	0.337	0.007	0.006	0.004	-	-	-	0.711	0.477	0.337
	RCNN	0.739	0.513	0.369	0.022	0.014	0.012	-	-	-	0.738	0.512	0.369
	CAML	0.727	0.522	0.378	0.018	0.012	0.011	0.004	0.003	0.003	0.726	0.521	0.377
	DR-CAML	0.713	0.517	0.374	0.023	0.014	0.011	0.005	0.004	0.003	0.712	0.517	0.373
	ZACNN	0.752	0.568	0.429	0.066	0.034	0.025	0.114	0.062	0.043	0.750	0.566	0.426
	ZAGCNN	0.778	0.596	0.454	0.077	0.043	0.030	0.118	0.063	0.046	0.777	0.595	0.453
	ACNN-KAMG($\mathbf{H}_g$)	0.777	0.588	0.450	0.084	0.046	0.032	0.160	0.070	0.048	0.776	0.587	0.450
	ACNN-KAMG($\mathbf{H}_s$)	0.772	0.593	0.451	0.090	0.047	0.034	0.143	0.074	0.052	0.770	0.592	0.450
	ACNN-KAMG($\mathbf{H}_c$)	0.772	0.594	0.454	0.088	0.048	0.034	0.141	0.071	0.050	0.770	0.593	0.453
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.770	0.590	0.451	0.086	0.044	0.031	0.155	0.075	0.050	0.769	0.589	0.450
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.769	0.590	0.450	0.087	0.046	0.032	0.153	0.071	0.049	0.768	0.589	0.449
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.784	0.599	0.458	0.109	0.054	0.037	0.186	0.080	0.054	0.783	0.598	0.457
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.785	0.597	0.456	0.100	0.053	0.038	0.146	0.077	0.054	0.784	0.596	0.456
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.785	0.595	0.455	0.085	0.047	0.033	0.159	0.081	0.055	0.783	0.594	0.454
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.780	0.595	0.453	0.092	0.051	0.035	0.165	0.081	0.056	0.778	0.594	0.452
MIMIC-III	CNN	0.826	0.684	0.548	0.020	0.012	0.009	-	-	-	0.826	0.684	0.548
	RCNN	0.845	0.702	0.560	0.034	0.021	0.016	-	-	-	0.845	0.701	0.560
	CAML	0.884	0.754	0.628	0.045	0.022	0.014	0.007	0.009	0.008	0.884	0.754	0.628
	DR-CAML	0.859	0.744	0.618	0.032	0.018	0.014	0.005	0.004	0.005	0.859	0.744	0.618
	ZACNN	0.858	0.728	0.603	0.010	0.035	0.026	0.007	0.069	0.052	0.858	0.710	0.567
	ZAGCNN	0.875	0.755	0.633	0.007	0.044	0.032	0.007	0.085	0.059	0.875	0.739	0.599
	ACNN-KAMG($\mathbf{H}_s$)	0.872	0.746	0.623	0.005	0.045	0.032	0.007	0.086	0.060	0.872	0.728	0.586
	ACNN-KAMG($\mathbf{H}_c$)	0.873	0.745	0.621	0.008	0.040	0.031	0.007	0.084	0.059	0.873	0.723	0.581
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.874	0.753	0.632	0.009	0.044	0.032	0.007	0.092	0.061	0.873	0.738	0.599
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.873	0.747	0.626	0.009	0.040	0.030	0.007	0.089	0.058	0.873	0.733	0.596
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.885	0.766	0.645	0.118	0.054	0.036	0.190	0.094	0.060	0.885	0.766	0.645
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.883	0.763	0.641	0.120	0.053	0.036	0.215	0.103	0.064	0.882	0.763	0.641
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.884	0.759	0.629	0.059	0.046	0.034	0.028	0.088	0.064	0.884	0.758	0.627
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.882	0.766	0.643	0.109	0.055	0.037	0.203	0.095	0.063	0.882	0.766	0.643
EU	AGRU-KAMG ($\mathbf{H}_g$)	0.857	0.581	0.361	0.415	0.158	0.094	0.247	0.103	0.060	0.862	0.596	0.375
	AGRU-KAMG ($\mathbf{H}_c$)	0.865	0.590	0.366	0.438	0.167	0.099	0.272	0.105	0.062	0.871	0.607	0.382
	AGRU-KAMG ($\mathbf{H}_s$)	0.866	0.588	0.366	0.447	0.168	0.099	0.276	0.105	0.064	0.873	0.625	0.382
	AGRU-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.881	0.606	0.373	0.496	0.184	0.107	0.285	0.108	0.065	0.889	0.626	0.393
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.882	0.606	0.373	0.489	0.183	0.107	0.276	0.105	0.062	0.888	0.626	0.392
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.884	0.607	0.375	0.491	0.184	0.107	0.323	0.112	0.069	0.891	0.627	0.394
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.883	0.610	0.376	0.504	0.188	0.110	0.290	0.115	0.068	0.891	0.630	0.396

Table 9: Precision@k results on MIMIC-II, MIMIC-III and EURLEX57K (EU) datasets

		Frequent			Few			Zero			Overall		
		RP@1	RP@5	RP@10	RP@1	RP@5	RP@10	RP@1	RP@5	RP@10	RP@1	RP@5	RP@10
MIMIC-II	CNN	0.712	0.478	0.337	0.007	0.006	0.004	-	-	-	0.711	0.477	0.337
	RCNN	0.739	0.513	0.369	0.022	0.014	0.012	-	-	-	0.738	0.512	0.369
	CAML	0.727	0.522	0.378	0.018	0.012	0.011	0.004	0.003	0.003	0.726	0.521	0.377
	DR-CAML	0.713	0.517	0.374	0.023	0.014	0.011	0.005	0.004	0.003	0.712	0.517	0.373
	ZACNN	0.752	0.568	0.429	0.066	0.034	0.025	0.114	0.062	0.043	0.750	0.566	0.426
	ZAGCNN	0.778	0.596	0.454	0.077	0.043	0.030	0.118	0.063	0.046	0.777	0.595	0.453
	ACNN-KAMG($\mathbf{H}_g$)	0.777	0.603	0.547	0.084	0.173	0.235	0.160	0.303	0.402	0.776	0.600	0.534
	ACNN-KAMG($\mathbf{H}_s$)	0.772	0.609	0.549	0.090	0.178	0.245	0.143	0.315	0.437	0.770	0.604	0.535
	ACNN-KAMG($\mathbf{H}_c$)	0.772	0.610	0.554	0.088	0.178	0.247	0.141	0.305	0.424	0.770	0.606	0.539
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.770	0.606	0.549	0.086	0.171	0.235	0.155	0.324	0.418	0.769	0.602	0.535
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.769	0.605	0.547	0.087	0.178	0.236	0.153	0.308	0.416	0.768	0.601	0.534
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.784	0.615	0.558	0.109	0.203	0.274	0.186	0.346	0.451	0.783	0.611	0.544
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.785	0.613	0.556	0.100	0.200	0.277	0.146	0.324	0.454	0.784	0.609	0.542
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.785	0.611	0.555	0.085	0.177	0.248	0.159	0.344	0.447	0.783	0.607	0.540
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.780	0.610	0.551	0.092	0.188	0.259	0.165	0.344	0.462	0.778	0.606	0.538
MIMIC-III	CNN	0.826	0.688	0.577	0.020	0.051	0.074	-	-	-	0.826	0.687	0.575
	RCNN	0.845	0.706	0.591	0.034	0.080	0.118	-	-	-	0.845	0.705	0.588
	CAML	0.884	0.759	0.662	0.045	0.084	0.104	0.007	0.036	0.067	0.884	0.758	0.659
	DR-CAML	0.859	0.749	0.652	0.032	0.073	0.105	0.005	0.016	0.038	0.859	0.749	0.649
	ZACNN	0.858	0.733	0.635	0.010	0.140	0.207	0.007	0.309	0.457	0.858	0.714	0.595
	ZAGCNN	0.875	0.759	0.668	0.007	0.181	0.258	0.007	0.367	0.512	0.875	0.743	0.629
	ACNN-KAMG($\mathbf{H}_s$)	0.872	0.750	0.657	0.005	0.184	0.258	0.007	0.376	0.524	0.872	0.732	0.615
	ACNN-KAMG($\mathbf{H}_c$)	0.873	0.750	0.656	0.008	0.171	0.252	0.007	0.374	0.518	0.873	0.727	0.610
	ACNN-KAMG ($\mathbf{H}_{g+s}$)	0.874	0.757	0.667	0.009	0.181	0.256	0.007	0.401	0.540	0.873	0.741	0.628
	ACNN-KAMG ($\mathbf{H}_{g+c}$)	0.873	0.752	0.661	0.009	0.167	0.245	0.007	0.397	0.521	0.873	0.737	0.625
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.885	0.771	0.680	0.118	0.224	0.293	0.190	0.412	0.530	0.885	0.770	0.677
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.883	0.768	0.676	0.120	0.217	0.284	0.215	0.449	0.560	0.882	0.768	0.673
	ACNN-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.884	0.763	0.663	0.059	0.192	0.280	0.028	0.386	0.566	0.884	0.762	0.658
	ACNN-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.882	0.770	0.679	0.109	0.231	0.295	0.203	0.417	0.553	0.882	0.770	0.675
EU	AGRU-KAMG ($\mathbf{H}_g$)	0.857	0.743	0.836	0.415	0.475	0.550	0.247	0.472	0.551	0.862	0.692	0.762
	AGRU-KAMG ($\mathbf{H}_c$)	0.865	0.755	0.847	0.444	0.504	0.585	0.272	0.488	0.574	0.871	0.705	0.775
	AGRU-KAMG ($\mathbf{H}_s$)	0.866	0.755	0.847	0.447	0.509	0.586	0.276	0.477	0.595	0.873	0.705	0.776
	AGRU-KAMG ($\mathbf{H}_c, \mathbf{H}_s$)	0.881	0.774	0.865	0.496	0.555	0.630	0.285	0.499	0.606	0.889	0.726	0.796
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s$)	0.882	0.778	0.858	0.489	0.551	0.631	0.276	0.480	0.569	0.888	0.734	0.795
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_c$)	0.884	0.776	0.868	0.491	0.555	0.628	0.323	0.517	0.641	0.891	0.728	0.799
	AGRU-KAMG ($\mathbf{H}_g, \mathbf{H}_s, \mathbf{H}_c$)	0.883	0.780	0.870	0.504	0.564	0.643	0.290	0.528	0.635	0.891	0.732	0.802

Table 10: R-Precision@k results on MIMIC-II, MIMIC-III and EURLEX57K (EU) datasets