# Graph-Driven Generative Models for Heterogeneous Multi-Task Learning

**Wenlin Wang**[1]**, Hongteng Xu**[2]**, Zhe Gan**[3]**, Bai Li**[1]**, Guoyin Wang**[1]
**Liqun Chen**[1]**, Qian Yang**[1]**, Wenqi Wang**[4]**, Ricardo Henao**[1]**, Lawrence Carin**[1]
[1]Duke University, [2]Infinia ML, [3]Microsoft Dynamics 365 AI Research, [4]Facebook
wenlin.wang@duke.edu

## Abstract

We propose a novel graph-driven generative model, that unifies multiple heterogeneous learning tasks into a unified framework. The proposed model is based on the fact that heterogeneous learning tasks, which correspond to different generative processes, often rely on data with a shared graph structure. Accordingly, our model combines a graph convolutional network (GCN) with multiple variational autoencoders (*i.e.*, graph-driven VAEs), thus embedding the nodes of the graph (*i.e.*, samples for the tasks) in a uniform manner while specializing their organization and usage to different tasks. With a focus on healthcare tasks, including clinical topic modeling, procedure recommendation and admission-type prediction, we show that our method successfully leverages information across different tasks, boosting performance in all tasks and outperforming existing state-of-the-art approaches.

## 1 Introduction

Heterogeneous multi-task learning aims to *jointly* solve different learning tasks, while each task potentially has a different objective. A central problem is to properly leverage information shared across all tasks [20, 4] and enrich the learning of each individual task. From the perspective of generative models, heterogeneous tasks usually correspond to distinct generative processes. This implies that traditional generative multi-task learning methods [2, 1, 26, 27] are not appropriate.

In this paper, we propose a graph-driven generative model to learn heterogeneous tasks in a unified framework. Taking advantage of the graph structure that commonly appears in many real-world data, the proposed model treats feature views, entities and their relationships as nodes and edges in a graph, and formulates learning heterogeneous tasks as instantiating different sub-graphs from the global data graph. Specifically, a sub-graph contains the feature views and the entities related to a task and their interactions. Both the feature views and the interactions can be reused across all tasks while the representation of the entities are specialized for the task. We combine a shared graph convolutional network (GCN) [10] with multiple variational autoencoders (VAEs) [9]. The GCN serves as a generator of latent representations for the sub-graphs, while the VAEs are specified to address the different tasks. The model is then optimized jointly over the objectives for all tasks to encourage the GCN to produce representations that can be used simultaneously by all of them.

We take health care as an motivating example, in which ICD (International Statistical Classification of Diseases) codes for diseases and procedures are used for multiple tasks, *e.g.*, modeling clinical topics of admissions, recommending procedures according to diseases and predicting admission types. In our work, ICD codes and hospital admissions (sets of ICD codes) constitute a graph as shown in Figure 1. The edges between ICD codes and those between ICD codes and admissions are quantified according to their coherency. The ICD codes are embedded during training, which are used to specialize the embeddings of admissions for different tasks. At test time, the GCN is used to represent sub-graphs, *i.e.*, collections of shared ICD codes, specialized admissions and their interactions, that feed into different task-specific VAEs. Experimental results show that the graph-driven framework indeed improves the performance of the three tasks described above.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

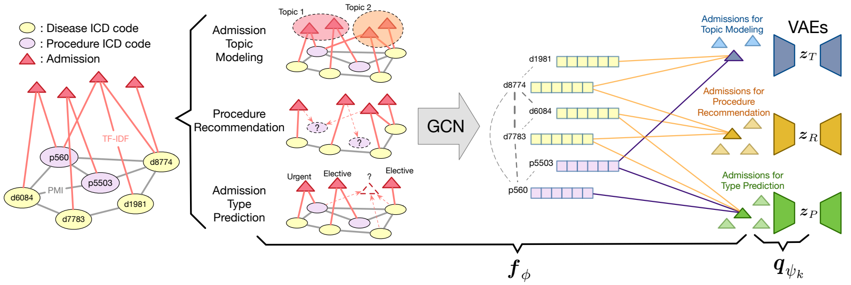

Figure 1: Illustration of the proposed model for healthcare tasks. Each task operates on a different sub-graph from the admission graph. The shared GCN ($f_\phi$) learns embeddings for ICD codes and admissions, and the embeddings pass through task-specific VAEs accordingly.

## 2 Proposed Model

A natural solution to solve heterogeneous multi-task learning from a generative model perspective is to model multiple generative processes, one for each task. In particular, given $K$ tasks, each task $k$ is associated with training data $(\boldsymbol{x}_k, \boldsymbol{y}_k)$, where $\boldsymbol{y}_k$ represents the target variable, and $\boldsymbol{x}_k$ represents the variable associated with $\boldsymbol{y}_k$. We propose using $K$ sets of VAEs [9] for modeling $\{\boldsymbol{y}_k\}_{k=1}^K$ in terms of latent variables $\{\boldsymbol{z}_k\}_{k=1}^K$, where each $\boldsymbol{z}_k$ is inferred from $\boldsymbol{x}_k$ using a task-specific inference network. The generative processes are defined as

$$\boldsymbol{y}_k \sim p_{\theta_k}(\boldsymbol{y}_k|\boldsymbol{z}_k)\,, \quad \boldsymbol{z}_k \sim p(\boldsymbol{z}_k)\,, \quad k = 1, \dots, K. \tag{1}$$

with corresponding inference networks specified as

$$\boldsymbol{z}_k \sim q_{\psi_k}(\boldsymbol{z}_k|f_\phi(\boldsymbol{x}_k))\,, \quad k = 1, \dots, K. \tag{2}$$

For the $k$-th task, $p_{\theta_k}(\cdot)$ represents a generative model, and $p(\boldsymbol{z}_k)$ is the prior distribution for latent code $\boldsymbol{z}_k$. The corresponding inference network for $\boldsymbol{z}_k$ consists of two parts: $(i)$ a deterministic encoder $f_\phi(\cdot)$ shared across all tasks to encode each $\boldsymbol{x}_k$ into $\hat{\boldsymbol{x}}_k = f_\phi(\boldsymbol{x}_k)$ independently; and $(ii)$ a stochastic encoder with parameters $\psi_k$ to stochastically map $\hat{\boldsymbol{x}}_k$ into latent code $\boldsymbol{z}_k$.

In likelihood-based learning, the goal for heterogeneous multi-task learning is to maximize the empirical expectation of the log-likelihood $\frac{1}{K}\sum_{k=1}^K \log(p(\boldsymbol{y}_k))$. Since the marginal likelihood $p(\boldsymbol{y}_k)$ rarely has a closed-form expression, VAE seeks to maximize the evidence lower bound (ELBO), which bounds the marginal log-likelihood as

$$\mathcal{L}(\theta_{1:K}, \psi_{1:K}, \phi) = \sum\nolimits_k \Big[\mathbb{E}_{q_{\psi_k}(\boldsymbol{z}_k|f_\phi(\boldsymbol{x}_k))}[\log p_{\theta_k}(\boldsymbol{y}_k|\boldsymbol{z}_k)] - \mathrm{KL}(q_{\psi_k}(\boldsymbol{z}_k|f_\phi(\boldsymbol{x}_k)) \parallel p(\boldsymbol{z}_k))\Big]\,. \tag{3}$$

However, for heterogeneous tasks, features are often organized in different views and the interactions between observed entities can as well be different. As a result, it is challenging to find a common $f_\phi(\cdot)$ for the $\{\boldsymbol{x}_k\}_{k=1}^K$ with incompatible formats or even in incomparable data spaces.

Fortunately, such data can often be modeled as a data graph, whose nodes correspond to the entities appearing in different tasks and edges capturing their complex interactions. Specifically, we represent a data graph as $G(\mathcal{V}, \mathcal{X}, \boldsymbol{A})$, where $\mathcal{V} = \{v_1, v_2, ...\}$ is the set of nodes corresponding to the observed entities, $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the adjacency matrix of the graph, and $\mathcal{X} = \cup_{k=1}^K \mathcal{X}_k$ is a union of (trainable) feature sets. Let $\mathcal{X}_k = \{\boldsymbol{x}_{v,k}\}_{v\in\mathcal{V}_k}$ be the feature set for the $k$-th task, where $\mathcal{V}_k \subset \mathcal{V}$ contains the nodes related to the task and $\boldsymbol{x}_{v,k}$ is the feature of the node $v$ in task $k$. Based on $\mathcal{X}_k$, the observations of the $k$-th task correspond to a sub-graph from $G$, *i.e.*, $G_k = G(\mathcal{V}_k, \mathcal{X}_k, \boldsymbol{A}_k)$, where $\boldsymbol{A}_k = \boldsymbol{A}(\mathcal{V}_k, \mathcal{V}_k)$ selects rows and columns from $\boldsymbol{A}$. To endow the framework a unified inference network, we implement $f_\phi(\cdot)$ in (3) with a graph convolutional network (GCN) [10] and thus $\boldsymbol{z}_k \sim q_{\psi_k}(\boldsymbol{z}_k|\mathrm{GCN}_\phi(G_k))$, with parameters of the inference network shared among tasks.

## 3 Typical Specification for Healthcare Tasks

To show the feasibility of our model, we describe a specification to solve tasks associated with hospital admissions. Let $\mathcal{V}^d = \{v_1^d, v_2^d, \dots\}$ and $\mathcal{V}^p = \{v_1^p, v_2^p, \dots\}$ denote the set of disease and procedure ICD codes, respectively, *i.e.*, each component $v_i^d \in \mathcal{V}^d$ represents a specific disease and each $v_i^p \in \mathcal{V}^p$ represents a specific procedure. Suppose we observe $N$ hospital admissions, denoted as $\mathcal{V}^a = \{v_1^a, v_2^a, \dots, v_N^a\}$. Each $v_n^a \in \mathcal{V}^a$ is associated with some ICD codes and a label representing its type, *i.e.*, $\{\mathcal{V}_n^d, \mathcal{V}_n^p, c_n\}$ for $n = 1, \dots, N$, where $\mathcal{V}_n^d \subseteq \mathcal{V}^d$, $\mathcal{V}_n^p \subseteq \mathcal{V}^p$ and $c_n \in \mathcal{C}$ is an element in the set of admission types $\mathcal{C}$. Based on these observations, we consider three healthcare tasks:

$i$) clinically-interpretable topic modeling of admissions; $ii$) procedure recommendation; and $iii$) admission-type prediction. A configuration of variables/graphs is highlighted in Table 5 in Appendix.

**Construction of edges** Inspired by existing research [12, 5, 16, 25], we enrich the representation power of our model with the meaningful population statistics. Two types of edges are considered.

($i$) *Edges between ICD codes.* ICD codes appear coherently in many admissions, *e.g.*, diabetes and its comorbidities like cardiovascular disease. Accordingly, edges between ICD codes with high coherency should be weighted heavily. Based on this principle, we apply point-wise mutual information (PMI) as the weight between each pair of ICD codes. Formally, for each pair of ICD codes, $\text{PMI}(i,j) = \log p_{ij} - \log(p_i p_j)$, for $i, j \in \mathcal{V}^d \cup \mathcal{V}^p$, where $p_{ij} = \frac{|\{v_n^a | i,j \in \mathcal{V}_n^d \cup \mathcal{V}_n^p\}|}{N}$ and $p_i = \frac{|\{v_n^a | i \in \mathcal{V}_n^d \cup \mathcal{V}_n^p\}|}{N}$. Positive PMI values indicate that the ICD codes in the pair are highly-correlated with each other. Conversely, negative PMI values imply weak correlation. Therefore, we only consider positive PMI values as the weights of edges.

($ii$) *Edges between ICD codes and admissions.* Analogous with the relationship between words and documents, we weight the edge between ICD codes and admissions with the term frequency-inverse document frequency (TF-IDF)[1], which defines how important an ICD code in an admission [15, 18].

Summarizing the above, elements $a_{ij}$ in the adjacency matrix $\boldsymbol{A}$ are represented as

$$a_{ij} = \begin{cases} 1, & i = j \\ \text{PMI}(i,j), & i, j \in \mathcal{V}^d \cup \mathcal{V}^p \text{ and } \text{PMI}(i,j) > 0 \\ \text{TF-IDF}(i,j), & i \in \mathcal{V}^a,\ j \in \mathcal{V}^d \cup \mathcal{V}^p \\ 0, & \text{otherwise} \end{cases} . \tag{4}$$

**Graph-driven VAEs for different tasks** We specify our model as graph-driven variational autoencoder (GD-VAE) and describe the aforementioned interested tasks as follows.

*(i) Topic modeling of admissions.* In the context of topic modeling, each ICD code can be considered as a *word* or *token*, while each admission corresponds to a *document*. However, patient admissions exhibit extreme-sparsity issues in the sense that a very small set of codes are associated with each admission. To circumvent this problem, inspired by [24], we model bi-term collections, and aggregate bi-terms from several admissions as one document. The generative process of our proposed Neural Bi-term Topic Model (NBTM) is as follows:

$$\boldsymbol{z}_T \sim \text{Dir}(\boldsymbol{\alpha}), \quad l \sim \text{Multi}(1, \boldsymbol{z}_T), \quad \boldsymbol{y}_T \sim \text{Multi}(2, \boldsymbol{\beta}_l)\,, \tag{5}$$

where $\boldsymbol{y}_T$ is the bi-term variable and its instance is a pair of ICD codes, $\{v_i, v_j\}$. $\boldsymbol{z}_T$ is the topic distribution. $\boldsymbol{\alpha}$ is the hyper-parameter of the Dirichlet prior; $\boldsymbol{\beta} = \{\boldsymbol{\beta}_l \in \mathbb{R}^{|\mathcal{V}^d|+|\mathcal{V}^p|}\}_{l=1}^{L}$ are trainable parameters, each representing a learned topic. The Dirichlet prior is known to be essential for generating interpretable topics [21] and it can be approximated with a multivariate logistic normal [19] for efficient inference.

($ii$) *Procedure recommendation.* For an admission, we aim to predict the set of procedures $\boldsymbol{y}_R$ for a set of diseases. Inspired by [11], we consider the following generative process:

$$\boldsymbol{z}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \pi_R \propto \exp\{g(\boldsymbol{z}_R)\}, \quad \boldsymbol{y}_R \sim \text{Multi}(M, \pi_R)\,, \tag{6}$$

where $\boldsymbol{y}_R$ is $|\mathcal{V}^p|$-dimensional variable and its instance is a list of $M$ recommended procedures. $g(\cdot)$ is a multi-layer perceptron (MLP). The output of this function is normalized to be a probability distribution over procedures, *i.e.*, $\pi_R \in \Delta^{|\mathcal{V}^p|}$, where $\Delta$ denotes a simplex. Then we derive procedures for the given admission by sampling $M$ times from a multinomial distribution with parameter $\pi_R$.

($iii$) *Admission-type prediction.* For an admission, the goal is to predict the admission type given both its diseases and procedures. We consider the generative process for modeling admission types as

$$\boldsymbol{z}_P \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \pi_P \propto \exp\{h(\boldsymbol{z}_P)\}, \quad \boldsymbol{y}_P \sim \text{Multi}(1, \pi_P)\,, \tag{7}$$

where $\boldsymbol{y}_P$ is a variable and its instance corresponds to an admission type in the set $\mathcal{C}$. $h(\cdot)$ is another MLP, The instance of $\boldsymbol{y}_R$ is sampled once from a multinomial distribution with parameter $\pi_P$.

**Inference with a shared GCN** The proposed model unifies three tasks via sharing a common GCN-based inference network. Specifically, the posteriors of the three latent variables are

$$\boldsymbol{z}_T \sim \mathcal{LN}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T),\ \boldsymbol{z}_R \sim \mathcal{N}(\boldsymbol{\mu}_R, \boldsymbol{\Sigma}_R),\ \boldsymbol{z}_P \sim \mathcal{N}(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P), \tag{8}$$

where $[\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k] = \text{MLP}_{\psi_k}(\text{GCN}_\phi(G_k))$ for $k \in \{P, R, T\}$. Let $\theta_T, \theta_R, \theta_P$ denote the parameters of the generative networks for each task, respectively. All the parameters $\{\theta_T, \theta_R, \theta_P, \psi_P, \psi_R, \psi_T, \phi\}$ are optimized jointly via maximizing (3).

[1] `https://en.wikipedia.org/wiki/Tf\T1\textendashidf`

# 4 Experiments

We test GD-VAE on a subset of MIMIC-III dataset [7], containing $31,213$ admissions with $2,765$ disease and 819 procedure ICD codes. The configuration of our experiments is presented in Appendix.

**Topic modeling.** Topic coherence [14] is used to evaluate the performance of topic modeling methods. Table 4 compares different methods on the mean of NPMI over the top 5/10/15/20 topic words. We find that LDA [3] performs worse than neural topic models (including ours), which demonstrates the necessity of introducing powerful inference networks to infer the latent topics. In terms of the GCN-based methods, GD-VAE and its variants capture global statistics between ICD codes and those between ICD codes and admissions, thus outperforming the three baselines by substantial margins. For leveraging knowledge across tasks, we find that the improvements are largely contributed by procedure recommendation, and marginally from admission prediction. This is because procedure recommendation accounts for the concurrence between disease codes and procedure codes within an admission, while the topic model considers the concurrence between the codes from different admissions. Both models capture the concurrence of ICD codes in different views, thus, naturally enhancing each other. We further visualize the top-5 ICD codes for some learned topics in the Appendix and find that the topic words are clinically-correlated.

| Method | T=10 | T=30 | T=50 |
|---|---|---|---|
| LDA [3] | 0.101 | 0.106 | 0.103 |
| AVITM [19] | 0.123 | 0.116 | 0.108 |
| BTM [24] | 0.104 | 0.110 | 0.107 |
| GD-VAE (T) | 0.128 | 0.129 | 0.123 |
| GD-VAE (TP) | 0.129 | 0.127 | 0.125 |
| GD-VAE (TR) | **0.136** | 0.133 | 0.127 |
| GD-VAE | **0.136** | **0.137** | **0.131** |

Table 1: Results on topic coherence for different models.

| Method | Top-1 (%) | | | Top-3 (%) | | | Top-5 (%) | | | Top-10 (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 | R | P | F1 |
| Word2Vec [13] | 5.3 | 22.9 | 8.7 | 14.6 | 21.1 | 15.3 | 24.8 | 21.0 | 20.1 | 41.1 | 17.7 | 22.2 |
| DWL [23] | 5.6 | 23.0 | 9.0 | 14.9 | 21.3 | 15.6 | 24.8 | 21.4 | 20.5 | 42.0 | 18.2 | 23.0 |
| BPR [17] | 7.3 | 26.7 | 10.2 | 23.0 | 27.1 | 21.2 | 38.4 | 27.6 | 27.9 | 56.6 | 21.7 | 28.0 |
| VAE-CF [11] | 17.8 | 50.1 | 23.5 | 35.2 | 37.9 | 33.4 | 47.9 | 32.4 | 34.6 | 63.0 | 21.7 | 30.2 |
| GD-VAE (R) | 20.1 | 53.4 | 25.8 | 37.2 | 40.1 | 35.5 | 49.1 | 32.5 | 35.2 | 64.6 | 23.7 | 31.0 |
| GD-VAE (RP) | 20.4 | 53.3 | 26.1 | 37.9 | 39.7 | 35.9 | 49.9 | 32.7 | 35.5 | 65.1 | 24.0 | 31.2 |
| GD-VAE (RT) | 20.9 | 56.2 | 27.2 | **41.0** | 42.2 | 36.5 | 50.9 | 35.1 | 36.6 | 66.0 | 24.7 | 32.5 |
| GD-VAE | **21.2** | **56.4** | **27.4** | 40.9 | **43.0** | **36.7** | **51.4** | **35.2** | **36.8** | **66.5** | **24.9** | **32.7** |

Table 2: Comparison of various methods on procedure recommendation.

**Procedure recommendation** Similar to [6, 23], we use top-$M$ precision, recall and F1-Score to evaluate the performance of procedure recommendation. Results are provided in Table 2. GD-VAE (R) is comparable to previous state-of-the-art algorithms. With additional knowledge learned from topic modeling and admission-type prediction, the results can be further improved. Similar to the observation in Section 4, topic modeling contributes more to procedure recommendation than admission-type prediction, since both topic modeling and procedure recommendation explore the underlying relationship between diseases and procedures.

**Admission-type prediction.** We employ precision, recall and F1-Score to evaluate the performance of admission-type prediction as well. Results in Table 3 show that GD-VAE outperforms its competitors. It is interesting to find that compared with topic modeling, procedure recommendation is more helpful to boost the results of admission-type prediction. One possible explanation is that the admission type is more relevant to the set of procedures, hence the embedding joint learned with procedure recommendation can better guide itself towards an accurate prediction, *e.g.,* it is likely to observe a *surgery* procedure in an *urgent* admission.

| Method | P | R | F1 |
|---|---|---|---|
| Word2Vec [13] | 87.11 | 89.16 | 88.12 |
| FastText [8] | 88.06 | 89.23 | 88.00 |
| SWEM [18] | 87.55 | 89.88 | 88.67 |
| LEAM [22] | 87.61 | 89.94 | 88.73 |
| GD-VAE (P) | 88.23 | 90.41 | 89.30 |
| GD-VAE (TP) | 88.31 | 90.56 | 89.41 |
| GD-VAE (RP) | 89.07 | 90.98 | 90.00 |
| GD-VAE | **89.14** | **91.01** | **90.05** |

Table 3: Results on admission-type prediction.

# 5 Conclusions

We have proposed a novel graph-driven variational autoencoder (GD-VAE) to learn multiple heterogeneous tasks within a unified framework. This is achieved by formulating entities under different tasks as different types of nodes, and using a shared GCN-based inference network to leverage knowledge across all tasks. Our model is general in that it can be easily extended to new tasks by specifying the corresponding generative processes. Experiments on real-world healthcare datasets demonstrate that GD-VAE can better leverage information across tasks, and achieve state-of-the-art results on clinical topic modeling, procedure recommendation, and admission-type prediction simultaneously.

## References

[1] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *JMLR*, 2003.

[2] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 1997.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[4] Rich Caruana. Multitask learning. *Machine learning*, 1997.

[5] Minmin Chen, Kilian Q Weinberger, Fei Sha, et al. An alternative text representation to tf-idf and bag-of-words. *CIKM*, 2012.

[6] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Sequential recommendation with user memory networks. In *WSDM*, 2018.

[7] 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. MIMIC-III, a freely accessible critical care database. *Scientific data*, 2016.

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv:1607.01759*, 2016.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[10] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[11] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *WWW*, 2018.

[12] Irina Matveeva. Document representation and multilevel measures of document similarity. In *NAACL*, 2006.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[14] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, 2011.

[15] Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 2016.

[16] Navid Rekabsaz, Bhaskar Mitra, Mihai Lupu, and Allan Hanbury. Toward incorporation of relevant documents in word2vec. *arXiv:1707.06598*, 2017.

[17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009.

[18] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *ACL*, 2018.

[19] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv:1703.01488*, 2017.

[20] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NIPS*, 1996.

[21] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *NIPS*, 2009.

[22] Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. *ACL*, 2018.

[23] Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled wasserstein learning for word embedding and topic modeling. In *NIPS*, 2018.

[24] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *WWW*, 2013.

[25] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *AAAI*, 2019.

[26] Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML*, 2005.

[27] Jian Zhang, Zoubin Ghahramani, and Yiming Yang. Flexible latent variable models for multi-task learning. *Machine Learning*, 2008.

|  | ICD codes | Description |
|---|---|---|
| Topic 1 | d8708 | Other Specified open wounds of ocular adnexa |
|  | d85306 | Other and unspecified intractranial hemorrhage following injury |
|  | dE8192 | Closed reduction of mandibular fracture |
|  | p7817 | Application of extrenal fixator device, tibia and fibula |
|  | p2751 | Suture of Iaceration of lip |
| Topic 2 | p3783 | Initial insertion of dual-chamber device |
|  | p3764 | Removal of external heart assist system or device |
|  | d7660 | Exceptionally large baby |
|  | 93514 | Open heart valvuloplasty of tricuspid value without replacement |
|  | d41021 | Acute myocardial infarction of inferolateral wall, initial episode of care |
| Topic 3 | d8774 | Retrograde pyelogram |
|  | d5503 | Percutaneous nephrostomy without gragmentation |
|  | d6084 | Other inflammatory disorders of male genital organs |
|  | p560 | Transurethral removal of obstruction from ureter and renal pelvis |
|  | d1981 | Secondary malignant neoplasm of other urinary organs |
| Topic 4 | p3951 | Clipping of aneurysm |
|  | d2242 | Frontal sunusectomy |
|  | p109 | Other cranial puncture |
|  | d78552 | Other craniotomy |
|  | d51883 | Other specified acquired deformity of head |
| Topic 5 | d33520 | Amyotrophic lateral sclerosis |
|  | d51902 | Mechanical complication of tracheostomy |
|  | p3199 | Other operations on trachea |
|  | d7708 | Other tracheostomy complications |
|  | d8718 | Chronic respiratory failure |
| Topic 6 | d7783 | Other hypothermia of newborn |
|  | p640 | Circumcision |
|  | d76406 | "light-for-dates" without mention of fetal malnutrition |
|  | d7731 | Hemolytic disease of fetus or newborn due to ABO isoimmunization |
|  | p9983 | Other phototherapy |
| Topic 7 | d45620 | Esophageal varices in diseases classified elsewhere, with bleeding |
|  | p9635 | Gastric Gavage |
|  | d4560 | Esophageal variaces with bleeding |
|  | d4233 | Endoscopic excision or destruction of lession or tissue of esophagus |
|  | d53240 | Chronic or unspecified duodenal ulcer with hemorrhage, without mention of obstruction |

Table 4: Full description of topic words.

# A Configurations of Our Method

We test various methods in 10 trials and record the mean value and standard deviation of the experimental results In each trial, we split the data into train, validation and test sets with a ratio of 0.6, 0.2 and 0.2, respectively. For the network architecture, we fix the embedding space to be 200 for ICD codes and admissions, and a two-layers GCN [10] with residual connection is considered for the inference network. In terms of the dimension of latent variable, $\boldsymbol{z}_T$ is identical to the number of topics for topic modeling and 200 for the other two tasks, $\boldsymbol{z}_R$ and $\boldsymbol{z}_P$. In the aspect of the generative network, a linear layer is employed for both topic modeling and admission type prediction. For the procedure recommendation, a one-hidden layer MLP with $\tanh$ as the nonlinear activation function is used. As for the hyper-parameters, we merge 10 randomly sampled admissions to generate a topic admission for our NBTM, such that $\boldsymbol{y}_T$ is not too sparse, and $5,000$ samples are generated so as to train the model. Following [19], the prior $\boldsymbol{\alpha}$ is a vector with constant value 0.02.

To investigate the affect of each components, we use "T", "R" and "P" to denote topic modeling, procedure recommendation and admission-type prediction, respectively. GD-VAE learns the three tasks jointly. To further verify the benefits of multi-task learning, we consider variations of our method that only learn one or two tasks, *e.g.*, GD-VAE (T) means only learning a topic model, and GD-VAE (TR) indicates the joint learning of topic modeling and procedure recommendation.

| Task | $G_k$ |  | $\boldsymbol{y}_k$ |
|---|---|---|---|
|  | $\mathcal{V}_k$ | $x^k_{v^a_n}$ in $\mathcal{X}_k$ |  |
| Topic Modeling | $\mathcal{V}$ | MaxPooling$(\{\boldsymbol{x}_v\}_{v\in\mathcal{V}^d_n\cup\mathcal{V}^p_n})$ | Bi-term ICD codes |
| Procedure Recommendation | $\mathcal{V}^d\cup\mathcal{V}^a$ | MaxPooling$(\{\boldsymbol{x}_v\}_{v\in\mathcal{V}^d_n})$ | List of procedures |
| Admission-type Prediction | $\mathcal{V}$ | MaxPooling$(\{\boldsymbol{x}_v\}_{v\in\mathcal{V}^d_n\cup\mathcal{V}^p_n})$ | Admission type, $c\in\mathcal{C}$ |

Table 5: Illustration of the differences between the three healthcare tasks.