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] Lawrence Carin[1]
[1]Duke University, [2]Infinia ML, Inc, [3]Microsoft Dynamics 365 AI Research, [4]Facebook, Inc
wenlin.wang@duke.edu

Abstract

We propose a novel graph-driven generative model, that unifies multiple heterogeneous learning tasks into the same 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, 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 applications (tasks), including clinical topic modeling, procedure recommendation and admission-type prediction, we demonstrate that our method successfully leverages information across different tasks, boosting performance in all tasks and outperforming existing state-of-the-art approaches.

Introduction

Multi-task learning aims to *jointly* solve different learning tasks, while leveraging appropriate information sharing across all tasks (Thrun 1996; Caruana 1997). It has been shown that learning under a multi-task setting usually yields enhanced performance relative to separately building single-task models (Sermanet et al. 2013; Hashimoto et al. 2016; Ruder 2017). However, multi-task learning has been considered primarily for *homogeneous* tasks that share the same objective (*e.g.*, the same set of labels) (Baxter 1997; Bakker and Heskes 2003; Yu, Tresp, and Schwaighofer 2005; Luo et al. 2015). Real-world tasks are often *heterogeneous* (Jin et al. 2014), meaning that each task potentially has a different objective and relies on complicated, often unobserved interactions. Examples of tasks having different objectives include classification, regression, recommendation, *etc*.

From the perspective of generative models, heterogeneous tasks often correspond to distinct generative processes. This implies that traditional generative multi-task learning methods (Baxter 1997; Bakker and Heskes 2003; Yu, Tresp, and Schwaighofer 2005; Zhang, Ghahramani, and Yang 2008), which often generalize a single class of generative model to multiple tasks, are not appropriate. Under these circumstances, a new mechanism is required to leverage relationships across the entities from different tasks. So motivated, 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) (Kipf and Welling 2017) with multiple variational autoencoders (VAEs) (Kingma and Welling 2013). 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.

In health care, our motivating example, ICD (International Statistical Classification of Diseases) codes for diseases and procedures can be used as a source of information for multiple tasks, *e.g.*, modeling clinical topics of admissions, recommending procedures according to diseases and predicting admission types. These three tasks require the capture of clinical relationships among ICD codes and admissions. A given admission has an associated set of disease and procedure codes (*i.e.*, feature views). However, the admission has to be organized with different views (*i.e.*, specialized entities) for tasks with different objectives. For instance, topic modeling is an unsupervised task needing procedures and diseases, admission-type prediction is a supervised task also using procedures and diseases, and procedure recommendation is a supervised task that only uses disease codes. In the context of our work, ICD codes and hospital admissions 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

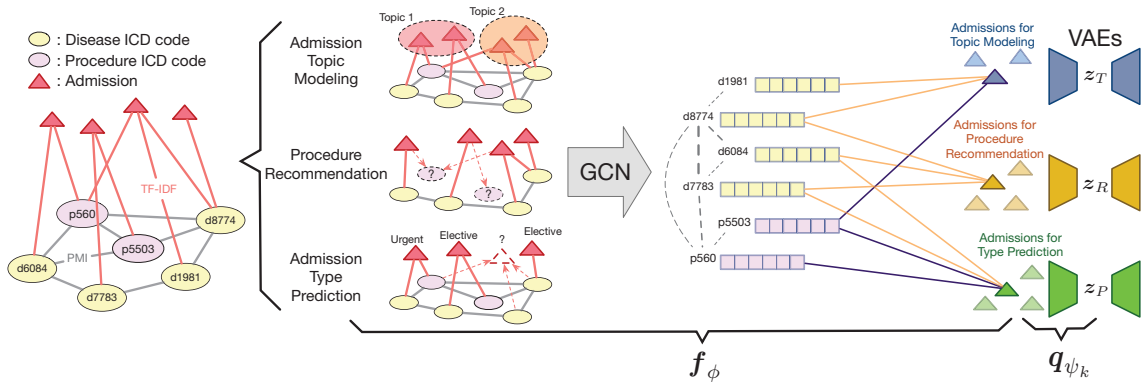


Figure 1: Illustration of the proposed model for healthcare tasks. Each task operates on a different sub-graph from the admission graph. The shared GCN (f_ϕ) learns embeddings for ICD codes and admissions, and the embeddings pass through task-specific VAEs.

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. We test our model on the three tasks described above. Experimental results show that the jointly learned representation for the admission graph indeed improves the performance of all tasks relative to the individual task model.

Proposed Model

To solve heterogeneous multi-task learning from a generative model perspective, a natural solution 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 (Kingma and Welling 2013) 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. Note that here the term VAE is used loosely in the sense that $\boldsymbol{y}_k$ and $\boldsymbol{x}_k$ need not to be the same. The generative processes are defined as

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

with corresponding inference networks specified as

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

For the k-th task, $p_{\theta_k}(\cdot)$ represents a generative model (*i.e.*, a stochastic decoder) with parameters θ_k, 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) an encoder with parameters ψ_k to stochastically map $\hat{\boldsymbol{x}}_k$ into latent code $\boldsymbol{z}_k$. The distribution $q_{\psi_k}(\boldsymbol{z}_k|f_\phi(\boldsymbol{x}_k))$ serves as an approximation to the unknown true posterior $p(\boldsymbol{z}_k|\boldsymbol{y}_k)$. Note that since $\{\boldsymbol{y}_k\}_{k=1}^K$ are in general associated with heterogeneous tasks, they may represent different types of information. For example, they can be labels for classification and bag-of-words for topic modeling. Motivated by the intuition that real-world tasks are likely to be related with each other in latent space, using a shared representation $f_\phi(\cdot)$ can be beneficial as a means to consolidate information in a way that allows tasks to leverage information from each other.

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

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

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 graph-based, with graph nodes corresponding to the entities appearing in different tasks and edges capturing their complex interactions. Accordingly, different tasks re-organize the graph and leverage its information from shared but different views. 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. For the k-th task, $\mathcal{X}_k = \{\boldsymbol{x}_{v,k}\}_{v\in\mathcal{V}_k}$ is its feature set, 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 $\boldsymbol{A}$. In such a situation, instead of finding a unified inference network for each individual observation in different tasks, for the sub-graphs we define an inference network based on a graph convolutional network (GCN) (Kipf and Welling 2017), *i.e.*,

implementing $f_\phi(\cdot)$ in (3) as a GCN with parameters ϕ and thus $\boldsymbol{z}_k \sim q_{\psi_k}(\boldsymbol{z}_k|\text{GCN}_\phi(G_k))$, hence a large portion of the parameters of the inference network are shared among tasks.

The independent generative processes with a shared GCN-based inference network match with the nature of heterogeneous tasks. In particular, the sub-graphs in different tasks are derived from the same data graph with partially shared nodes and edges, enabling joint learning of latent variables through the shared inference network. Then, the inferred latent variables pass through different generative models under the guidance of different tasks. In the next section, we demonstrate that this model is suitable for challenging healthcare tasks.

Typical Specification for Healthcare Tasks

Observations, tasks, and proposed data graph

To demonstrate the feasibility of our model, we describe a specification to solve tasks associated with hospital admissions. Let $\mathcal{V}^d = \{v_1^d, v_2^d, \ldots\}$ and $\mathcal{V}^p = \{v_1^p, v_2^p, \ldots\}$ 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 $\mathcal{V}^a = \{v_1^a, v_2^a, \ldots, v_N^a\}$. Each $v_n^a \in \mathcal{V}^a$ is associated with ICD codes and a label representing its type, *i.e.*, $\{\mathcal{V}_n^d, \mathcal{V}_n^p, c_n\}$ for $n = 1, \ldots, N$, 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.

As illustrated in Figure 1, the observations can be represented as an admission graph $G(\mathcal{V}, \mathcal{X}, \boldsymbol{A})$, where the node set $\mathcal{V} = \mathcal{V}^d \cup \mathcal{V}^p \cup \mathcal{V}^a$ and $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the adjacency matrix. The union of feature sets $\mathcal{X} = \mathcal{X}^d \cup \mathcal{X}^p \cup \mathcal{X}^a$, where $\mathcal{X}^d = \{\boldsymbol{x}_v\}_{v\in\mathcal{V}^d}$ and $\mathcal{X}^p = \{\boldsymbol{x}_v\}_{v\in\mathcal{V}^p}$ contain trainable vector embeddings of ICD codes for diseases and procedures, respectively. These embeddings are reused for different tasks. $\mathcal{X}^a = \cup_{k=1}^K \mathcal{X}_k^a$, where $\mathcal{X}_k^a = \{\boldsymbol{x}_v^k\}_{v\in\mathcal{V}^a}$ contains the embeddings of admissions for different tasks. Specifically, for each admission v_n^a, its embedding in the k-th task is derived from the aggregation of the embeddings of ICD codes, *i.e.*, $\boldsymbol{x}_{v_n^a}^k = \text{MaxPooling}(\{\boldsymbol{x}_v\}_{v\in\mathcal{V}_n^k})$, where $\mathcal{V}_n^k$ is the set of the ICD codes associated with task k. For topic modeling and admission-type prediction, $\mathcal{V}_n^k = \mathcal{V}_n^d \cup \mathcal{V}_n^p$, while for procedure recommendation, in which the procedure codes are unavailable, $\mathcal{V}_n^k = \mathcal{V}_n^d$. Given admission graph G, the three healthcare tasks correspond to different sub-graphs $G(\mathcal{V}_k, \mathcal{X}_k, \boldsymbol{A}_k)$, which yield a typical heterogeneous scenario. Table 1 highlights their differences on target variables and sub-graphs. Although the sub-graphs specialize the information of admission nodes, they reuse the representations of ICD code nodes and the edges in G.

Construction of edges

Inspired by existing research (Matveeva 2006; Chen et al. 2012; Rekabsaz et al. 2017; Yao, Mao, and Luo 2019), we enrich the representation power of our model with meaningful population statistics, considering two types of edges in our adjacency matrix.

Edges between ICD codes. ICD codes are correlated 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), which is a commonly-used similarity measurement in various NLP tasks (Levy and Goldberg 2014; Arora et al. 2016; Newman et al. 2010; Mimno et al. 2011; Ogura and Kobayashi 2013), as the weight between each pair of ICD codes. Formally, for each pair of ICD codes, we evaluate the PMI as

$$\text{PMI}(i,j) = \log p_{ij} - \log(p_i p_j), \text{ for } i,j \in \mathcal{V}^d \cup \mathcal{V}^p, \quad (4)$$

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.

Edges between ICD codes and admissions. Analogous to the relationship between words and documents, we weight the edge between ICD codes and admissions with the help of the term frequency-inverse document frequency (TF-IDF)[1]. The term frequency (TF) is the normalized version of the number of times an ICD code appears in an admission, and the inverse document frequency (IDF) is the log-scaled inverse fraction of the number of admissions that contain the ICD code. The TF-IDF is the element-wise multiplication of TF and IDF, which defines how important an ICD code is in an admission (Onan, Korukoğlu, and Bulut 2016; Shen et al. 2018).

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

$$a_{ij} = \begin{cases} 1, & i = j \\ \text{PMI}(i,j), & i,j \in \mathcal{V}^d \cup \mathcal{V}^p \text{ and } \text{PMI}(i,j) > 0 \\ \text{TF-IDF}(i,j), & i \in \mathcal{V}^a,\ j \in \mathcal{V}^d \cup \mathcal{V}^p \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

Graph-driven VAEs for different tasks

Focusing on the three tasks mentioned above, we specify our model as graph-driven variational autoencoder (GD-VAE). Specifically, GD-VAE consist of: i) a GCN-based inference network that is shared by all the three tasks, and ii) three specialized generative networks that account for different sets of observations corresponding to the three tasks.

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*, *i.e.*, a collection of ICD codes. However, patient admissions exhibit extreme-sparsity issues in the sense that a very small set of codes are associated with each admission. Classic topic models, such as latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) and the Neural Topic Model (Miao, Grefenstette, and Blunsom 2017), can therefore be inappropriate in this case. To circumvent this, inspired by (Yan et al.

[1] https://en.wikipedia.org/wiki/Tf-idf

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 1: Illustration of the differences between the three healthcare tasks.

2013), instead of modeling a bag-of-ICD-codes for a single admission, we aim to model bi-term collections, and then aggregate all the unordered ICD code pairs (bi-terms) from several admissions together as one document. The generative process of our proposed Neural Bi-term Topic Model (NBTM) is described as follows:

$$\boldsymbol{z}_T \sim \text{Dir}(\boldsymbol{\alpha}), \quad l \sim \text{Multi}(1, \boldsymbol{z}_T), \quad \boldsymbol{y}_T \sim \text{Multi}(2, \boldsymbol{\beta}_l)\,, \tag{6}$$

where $\boldsymbol{y}_T$ is the bi-term variable and its instance is $\{v_i, v_j\}$, where v_i, v_j are two ICD codes. $\boldsymbol{z}_T$ is the topic distribution. $\boldsymbol{\alpha}$ is the hyper-parameter of the Dirichlet prior, a vector with length L, where L is the number of topics. $\boldsymbol{\beta} = \{\boldsymbol{\beta}_l \in \mathbb{R}^{|\mathcal{V}^d|+|\mathcal{V}^p|}\}_{l=1}^L$ are trainable parameters, each representing a learned topic, *i.e.*, the distribution over ICD codes. The marginal likelihood for the entire admission corpus $\boldsymbol{y}_T$ can be written as

$$\begin{aligned} p(\boldsymbol{y}_T) &= \int_{\boldsymbol{z}_T} p(\boldsymbol{z}_T) \prod\nolimits_{i,j} \sum\nolimits_l p(v_i, v_j|\boldsymbol{\beta}_l) p(l|\boldsymbol{z}_T) d\boldsymbol{z}_T \\ &= \int_{\boldsymbol{z}_T} p(\boldsymbol{z}_T) \prod\nolimits_{i,j} p(v_i, v_j|\boldsymbol{\beta}, \boldsymbol{z}_T) d\boldsymbol{z}_T\,. \end{aligned} \tag{7}$$

The Dirichlet prior is known to be essential for generating interpretable topics (Wallach, Mimno, and McCallum 2009). However, it can rarely be applied to a VAE directly, since no effective re-parameterization can be adopted for the Dirichlet distribution. Fortunately, the Dirichlet distribution can be approximated with a logistic normal and a softmax formulation by the Laplace approximation (Hennig et al. 2012). When the number of topics L is large, the Dirichlet distribution can be approximated with a multivariate logistic normal (Srivastava and Sutton 2017) with the i-th element of its mean $\boldsymbol{\mu}_T$ and diagonal covariance matrix $\boldsymbol{\Sigma}_T$ as follows:

$$\begin{aligned} \mu_i^T &= \log \alpha_i - \frac{1}{L}\sum\nolimits_{j=1}^L \log \alpha_j, \\ \Sigma_{ii}^T &= \frac{1}{\alpha_i}(1 - \frac{2}{L}) + \frac{1}{L^2}\sum\nolimits_{j=1}^L \frac{1}{\alpha_j}\,. \end{aligned} \tag{8}$$

Under such an approximation, a topic distribution can be readily inferred by applying re-parameterization, sampling $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and inferring $\boldsymbol{z}_T$ via $\boldsymbol{z}_T = \text{softmax}(\boldsymbol{\mu}_T + \boldsymbol{\Sigma}_T^{1/2}\boldsymbol{\epsilon})$.

Procedure recommendation In this task, for an admission, we aim to predict the set of procedures $\boldsymbol{y}_R$ for a set of diseases. Inspired by (Liang et al. 2018), we consider the following generative process for modeling admission procedures:

$$\boldsymbol{z}_R \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \pi_R \propto \exp\{g(\boldsymbol{z}_R)\}, \quad \boldsymbol{y}_R \sim \text{Multi}(M, \pi_R)\,, \tag{9}$$

	Small	**Median**	**Large**
$\lvert\mathcal{V}^d\rvert$	247	874	2,765
$\lvert\mathcal{V}^p\rvert$	75	258	819
$\lvert\mathcal{V}^a\rvert$	28,315	30,535	31,213

Table 2: Statistics of the MIMIC-III dataset.

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 Δ denotes a simplex. Then we derive procedures for the given admission by sampling M times from a multinomial distribution with parameter π_R.

Admission-type prediction Given an admission, the goal is to predict the admission type given both its diseases and procedures. We consider the following generative process for modeling admission types:

$$\boldsymbol{z}_P \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \pi_P \propto \exp\{h(\boldsymbol{z}_P)\}, \quad \boldsymbol{y}_P \sim \text{Multi}(1, \pi_P)\,, \tag{10}$$

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, whose output is normalized to be a distribution over admission types, *i.e.*, $\pi_P \in \Delta^{|\mathcal{C}|}$. Finally, the instance of $\boldsymbol{y}_R$ (the type of the given admission) is sampled once from a multinomial distribution with parameter π_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

$$\boldsymbol{z}_T \sim \mathcal{LN}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T),\ \boldsymbol{z}_R \sim \mathcal{N}(\boldsymbol{\mu}_R, \boldsymbol{\Sigma}_R),\ \boldsymbol{z}_P \sim \mathcal{N}(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P), \tag{11}$$

where $\text{diag}(\boldsymbol{\sigma}_k) = \boldsymbol{\Sigma}_k^{1/2}$, $\text{diag}(\cdot)$ represents a diagonal matrix, and $[\boldsymbol{\mu}_k; \log \boldsymbol{\sigma}_k] = \text{MLP}_{\psi_k}(\text{GCN}_\phi(G_k))$ for $k \in \{P, R, T\}$.

Let $\theta_T, \theta_R, \theta_P$ denote the parameters of the generative networks for topic modeling, procedure recommendation and admission-type prediction, respectively. All parameters $\{\theta_T, \theta_R, \theta_P, \psi_P, \psi_R, \psi_T, \phi\}$ are optimized jointly via maximizing (3).

Related Work

Multi-task learning Many multi-task learning methods learn a shared latent representation (Thrun 1996; Caruana 1997; Baxter 1997), or impose structural constraints on the shared features for different tasks (Ando and Zhang 2005;

Method	Small		Median	Large
	T=10T=30	T=50	T=10T=30T=50	T=10T=30T=50
LDA (Blei, Ng, and Jordan 2003)	0.1100.106	0.098	0.1230.1020.107	0.1010.1060.103
AVITM (Srivastava and Sutton 2017)	0.1320.125	0.121	0.1350.1100.107	0.1230.1160.108
BTM (Yan et al. 2013)	0.1170.109	0.105	0.1270.1080.105	0.1040.1100.107
GD-VAE (T)	0.1420.141	0.135	0.1400.1370.132	0.1280.1290.123
GD-VAE (TP)	0.1420.138	0.136	0.1430.1370.134	0.1290.1270.125
GD-VAE (TR)	0.1470.147	0.144	0.1460.1410.137	**0.136**0.1330.127
GD-VAE	**0.1510.149**	**0.145**	**0.1480.1440.140**	**0.1360.1370.131**

Standard deviation for GD-VAE and its variants is around 0.003.

Table 3: Results on topic coherence for different models.

Chen et al. 2009). The work in (He and Lawrence 2011) proposed a graph-based framework leveraging information across multiple tasks and multiple feature views. Following this work, the methods in (Zhang and Huan 2012; Jin et al. 2013) applied structural regularizers across different feature views and tasks, and ensured the learned predictive models are similar for different tasks. However, these methods require that multiple tasks directly share some label-dependent information with each other, which is only applicable to homogeneous tasks. Focusing on heterogeneous tasks, many discriminative methods have been proposed, which map the original heterogeneous features to a shared latent space, through linear or nonlinear functions (Zhang and Yeung 2011; Jin et al. 2014; Liu et al. 2018b) or sparsity-driven feature selection (Yang, Kim, and Xing 2009; Jin et al. 2015), and solve heterogeneous tasks jointly in the framework of discriminant analysis. Generative models have achieved success in the past few years (Wang et al. 2017; 2018b; 2019). However, to our knowledge, the generative solutions to heterogeneous multi-task learning have not been fully investigated.

ICD code embedding and analysis of healthcare data Machine learning techniques have shown potential in many healthcare problems, *e.g.*, ICD code assignment (Shi et al. 2017; Baumel et al. 2017; Mullenbach et al. 2018; Huang, Osorio, and Sy 2018), admission prediction (Ma et al. 2017; Liu et al. 2018a; Xu et al. 2017), mortality prediction (Harutyunyan et al. 2017; Xu et al. 2018), procedure recommendation (Mao, Yao, and Luo 2019), medical topic modeling (Choi et al. 2017; Suo et al. 2018), *etc.* Although these tasks have different objectives, they often share the same electronic health records data, *e.g.*, admission records. To learn multiple healthcare tasks jointly, various multi-task learning methods have been proposed (Wang, Wang, and Hu 2014; Alaa and van der Schaar 2017; Suo et al. 2017; Harutyunyan et al. 2017; Mao, Yao, and Luo 2019). Traditional multi-task learning methods often impose structural regularizers on the features shared by different tasks (Argyriou, Evgeniou, and Pontil 2007). The work in (Mao, Yao, and Luo 2019) applied GCNs (Kipf and Welling 2017) to extract features and jointly train models for medication recommendation and lab-test imputation, which constitutes an attempt to apply GCNs to multi-task learning. However, introducing GCNs into the framework of generative heterogenous multi-task learning remains unexplored, that this paper seeks to address.

Experiments

We test our method (GD-VAE) on the MIMIC-III dataset (Johnson et al. 2016), which contains more than 58,000 hospital admissions with 14,567 disease ICD codes and 3,882 procedures ICD codes. For each admission, it consists of a set of disease and procedure ICD codes. Three subsets of the MIMIC-III data are considered, with summary statistics in Table 2. The subsets are generated by thresholding the frequency of ICD codes, *i.e.*, the ICD codes appearing at least 500/100/50 times and the corresponding non-empty admissions constitute the small/median/large subset.

To demonstrate the effectiveness of our method, we compare GD-VAE with state-of-the-art approaches on each of the healthcare tasks mentioned above. Specifically, (i) for topic modeling, we compare with LDA (Blei, Ng, and Jordan 2003), AVITM (Srivastava and Sutton 2017) and BTM (Yan et al. 2013). (ii) For procedure recommendation, we compare with Bayesian Personalized Ranking (BPR) (Rendle et al. 2009), Distilled Wasserstein Learning (DWL) (Xu et al. 2018), and a VAE model designed for collaborative filtering (VAE-CF) (Liang et al. 2018). We also compare with a baseline method based on Word2Vec (Mikolov et al. 2013), which enumerates all possible disease-procedure pairs in each admission, and then recommends procedures according to the similarity between their embeddings and those of diseases. (iii) For admission-type prediction, we consider the following baselines: TF-IDF (combined with a linear classifier), Word2Vec (learning ICD code embeddings with Word2Vec (Mikolov et al. 2013), and using the mean of the learned embeddings to predict the label), FastText (Joulin et al. 2016), SWEM (Shen et al. 2018) and LEAM (Wang et al. 2018a). 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.

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

Dataset	Method	Top-1 (%)			Top-3 (%)			Top-5 (%)			Top-10 (%)		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1
Small	Word2Vec (Mikolov et al. 2013)	19.5	47.8	24.7	35.4	34.9	30.8	47.1	29.6	32.0	62.3	21.2	8.5
	DWL (Xu et al. 2018)	19.7	48.2	25.0	35.9	35.2	31.3	47.5	30.3	32.4	63.0	20.9	28.7
	BPR (Rendle et al. 2009)	23.5	57.6	29.8	44.8	43.5	38.7	56.8	35.7	38.8	73.1	24.8	33.6
	VAE-CF (Liang et al. 2018)	24.0	57.8	30.7	46.0	43.5	39.3	57.8	35.2	39.1	74.0	24.2	33.8
	GD-VAE (R)	24.8	58.2	31.1	46.5	43.4	39.5	58.1	35.3	39.2	74.5	24.4	34.0
	GD-VAE (RP)	25.0	58.3	31.3	46.8	43.5	39.5	58.2	35.4	39.2	74.7	24.5	34.1
	GD-VAE (RT)	25.4	58.3	31.6	**47.0**	43.6	39.7	58.5	35.9	39.4	75.2	24.8	34.3
	GD-VAE	**25.6**	**58.6**	**31.8**	**47.0**	**43.8**	**39.8**	**58.7**	**36.2**	**39.6**	**75.9**	**25.1**	**34.5**
Median	Word2Vec (Mikolov et al. 2013)	7.8	27.6	10.9	27.7	30.5	25.1	38.3	26.9	27.7	52.8	20.1	26.1
	DWL (Xu et al. 2018)	8.0	27.5	11.1	27.9	30.8	25.2	39.5	27.0	27.9	53.9	20.9	27.4
	BPR (Rendle et al. 2009)	10.2	35.8	14.9	38.6	40.2	34.3	49.3	33.3	34.9	65.2	23.8	31.4
	VAE-CF (Liang et al. 2018)	21.2	52.9	26.2	41.2	42.0	36.0	53.4	35.3	37.3	68.2	24.9	32.9
	GD-VAE (R)	22.0	55.1	27.9	42.3	41.2	37.2	54.0	35.7	37.8	69.3	25.2	33.1
	GD-VAE (RP)	22.3	55.1	28.0	42.7	41.5	37.4	53.7	35.5	37.6	69.6	25.1	33.4
	GD-VAE (RT)	22.8	57.8	29.3	43.0	43.5	38.1	54.2	35.9	38.1	70.1	25.2	33.6
	GD-VAE	**23.2**	**57.9**	**29.6**	**43.2**	**43.9**	**38.2**	**54.6**	**36.0**	**38.4**	**70.4**	**25.3**	**33.7**
Large	Word2Vec (Mikolov et al. 2013)	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 (Xu et al. 2018)	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 (Rendle et al. 2009)	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 (Liang et al. 2018)	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	23.1	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**

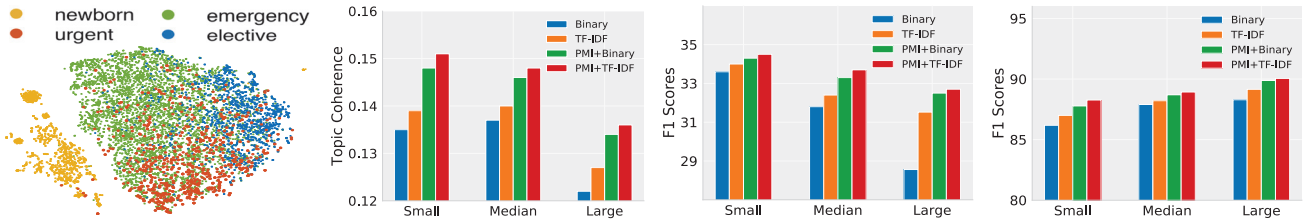
The standard deviation for GD-VAE and its variants is less than 0.2.

Table 4: Comparison of various methods on procedure recommendation.

Data	Small			Median			Large		
Method	P	R	F1	P	R	F1	P	R	F1
TF-IDF	84.2	87.1	85.18	86.1	28.6	87.22	88.4	58.1	087.76
Word2Vec (Mikolov et al. 2013)	85.0	88.7	86.23	86.6	088.8	787.71	87.1	189.1	688.12
FastText (Joulin et al. 2016)	84.2	187.1	585.29	86.6	688.6	587.39	88.0	689.2	388.00
SWEM (Shen et al. 2018)	85.5	688.1	086.77	87.0	189.2	888.12	87.5	589.8	888.67
LEAM (Wang et al. 2018a)	85.3	488.0	386.55	87.0	389.2	988.14	87.6	189.9	488.73
GD-VAE (P)	86.0	188.1	386.91	87.7	689.3	188.51	88.2	390.4	189.30
GD-VAE (TP)	86.1	888.5	287.22	87.8	289.2	188.52	88.3	190.5	689.41
GD-VAE (RP)	86.8	789.3	887.93	88.0	889.5	788.82	89.0	790.9	890.00
GD-VAE	**87.0**	**089.6**	**088.01**	**88.1**	**989.7**	**088.94**	**89.1**	**491.0**	**190.05**

The standard deviation for GD-VAE and its variants is around 0.05 on F1 score.

Table 5: Results on admission-type prediction.



(a) t-SNE visualization (b) Topic Modeling (c) Procedure Recommendation (d) Admission-type Prediction

Figure 2: (a) t-SNE visualization of $\boldsymbol{z}_P$. (b, c, d) Rationality analysis of graph construction.

ICD codes and admissions, and a two-layer GCN (Kipf and Welling 2017) with residual connection is considered for the inference network. Concerning 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 $\boldsymbol{z}_P$. For 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. Concerning the hyperparameters, 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 (Srivastava and Sutton 2017), the prior $\boldsymbol{\alpha}$ is a vector with constant value 0.02.

Topic modeling

Topic coherence (Mimno et al. 2011) is used to evaluate the performance of topic modeling methods. This metric is computed based on the normalized point-wise mutual information (NPMI), which has proven to match well with human judgment (Lau, Newman, and Baldwin 2014). Table 3 compares different methods on the mean of NPMI over the top

Table 6: Full description of topic words.

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

5/10/15/20 topic words. We find that LDA (Blei, Ng, and Jordan 2003) 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.

Compared with only performing topic modeling, *i.e.*, GD-VAE (T), considering more tasks brings improvements, and the proposed GD-VAE achieves the best performance. Concerning 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.

To further verify the quality of the learned topics, we visualize the top-5 ICD codes for some learned topics in the Supplementary Material. We find that the topic words are clinically-correlated. For example, the ICD codes related to *surgery* and those about *urology* are concentrated in two topics. Additionally, each topic contains both disease codes and procedure codes, *e.g.,* "d85306" and "p7817" are *orthopedic surgery* related disease and procedures, showing that disease and procedures can be closely correlated, which also implies the potential benefits brought to procedure recommendation.

Procedure recommendation

Similar to (Chen et al. 2018; Xu et al. 2018), we use top-M precision, recall and F1-Score to evaluate the performance of procedure recommendation. Given the n-th admission, we denote W_n and Y_n as the top-M list of recommended procedures and ground-truth procedures, respectively. The top-M precision, recall and F1-score are calculated as follows: $P = \sum_{n=1}^{N} P_i = \sum_{n=1}^{N} \frac{|W_n \cap Y_n|}{|W_n|}$, $R = \sum_{n=1}^{N} R_i = \sum_{i=1}^{N} \frac{|W_n \cap Y_n|}{|Y_n|}$, $F1 = \sum_{n=1}^{N} \frac{2P_nR_n}{P_n+R_n}$. Results are provided in Table 4. 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 the previous section, 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

Similar to procedure recommendation, we use precision, recall and F1-Score to evaluate the performance of admission-type prediction. Results in Table 5 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 jointly 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. Additionally, to better understand the representation learned by GD-VAE, we visualize the inferred latent code $\boldsymbol{z}_P$ with t-SNE (Maaten and Hinton 2008), as shown in Figure 2a. The $\boldsymbol{z}_P$ is visually representative of different admission types being clustered into different groups.

Rationality of graph construction

To explore the rationality of our graph construction, we also compare the proposed admission graph with its variants. In particular, the proposed admission graph considers the PMI between ICD codes and the TF-IDF between ICD codes and admissions (*i.e.*, PMI + TF-IDF). Its variations include: (i) a simple graph with binary edges between admissions and ICD codes (Binary); (ii) a graph only considering the TF-IDF between admissions and ICD codes (TF-IDF); and (iii) a graph considering the PMI between ICD codes and the binary edges between admissions and ICD codes (PMI + Binary). Results using different graphs are provided in Figures 2b through 2d, which demonstrate that both PMI edges and TF-IDF edges make significant contributions to the performance of the proposed GD-VAE.

Conclusions

We have proposed a 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. Comprehensive experiments on real-world healthcare data 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.

Acknowledgement: This research was supported in part by DARPA, DOE, NIH, ONR and NSF.

GCN-based Inference Network

Graph convolutional network (GCN) (Kipf and Welling 2017) has attracted much attention for leveraging information representations for nodes and edges, and is promising to tasks with complex relational information (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017; Yao, Mao, and Luo 2019).

Given a graph $G(\mathcal{V}, \mathcal{X}, \boldsymbol{A})$, a graph convolution layer derives the node embeddings $\boldsymbol{H} \in \mathbb{R}^{N \times \tilde{d}}$ via

$$\boldsymbol{H} = \mathrm{ReLU}(\tilde{\boldsymbol{A}}\mathcal{X}\boldsymbol{W}),$$

where $\tilde{d}$ is the dimension of feature space after GCN, $\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ is the normalized version of adjacency matrix, $\boldsymbol{D}$ is a diagonal degree matrix with $\boldsymbol{D}_{ii} = \sum_j \tilde{\boldsymbol{A}}_{ij}$, and $\boldsymbol{W}$ is a matrix of trainable graph convolution parameters. The GCN aggregates node information in local neighborhoods to extract local substructure information. In order to incorporate high-order neighborhoods, we can stack multiple graph convolution layers as

$$\boldsymbol{H}^{l+1} = \mathrm{ReLU}(\tilde{\boldsymbol{A}}\boldsymbol{H}^l\boldsymbol{W}^l),$$

where $\boldsymbol{H}^0 = \boldsymbol{X}$, $\boldsymbol{H}^l$ is the output of the l-th graph convolutional layer. However, GCN can be interpreted as Laplacian smoothing. Repeatedly applying Laplacian smoothing may fuse the features over vertices and make them indistinguishable (Li, Han, and Wu 2018). Inspired by (He et al. 2016), we alleviate this problem in our inference network by adding shortcut connections between different layers.

Description of topic ICD codes

A full description of the top-5 topic ICD codes is shown in Table 6.

References

Alaa, A. M., and van der Schaar, M. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *NIPS*.

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In *NIPS*.

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2016. A latent variable model approach to pmi-based word embeddings. *ACL*.

Bakker, B., and Heskes, T. 2003. Task clustering and gating for bayesian multitask learning. *JMLR*.

Baumel, T.; Nassour-Kassis, J.; Elhadad, M.; and Elhadad, N. 2017. Multi-label classification of patient notes a case study on ICD code assignment. *arXiv:1709.09587*.

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

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR*.

Caruana, R. 1997. Multitask learning. *Machine learning*.

Chen, J.; Tang, L.; Liu, J.; and Ye, J. 2009. A convex formulation for learning shared structures from multiple tasks. In *ICML*.

Chen, M.; Weinberger, K. Q.; Sha, F.; et al. 2012. An alternative text representation to tf-idf and bag-of-words. *CIKM*.

Chen, X.; Xu, H.; Zhang, Y.; Tang, J.; Cao, Y.; Qin, Z.; and Zha, H. 2018. Sequential recommendation with user memory networks. In *WSDM*.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. Gram: graph-based attention model for healthcare representation learning. In *KDD*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NIPS*.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; and Galstyan, A. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv:1703.07771*.

Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv:1611.01587*.

He, J., and Lawrence, R. 2011. A graph-based framework for multi-task multi-view learning. In *ICML*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hennig, P.; Stern, D.; Herbrich, R.; and Graepel, T. 2012. Kernel topic models. In *AISTATS*.

Huang, J.; Osorio, C.; and Sy, L. W. 2018. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *arXiv:1802.02311*.

Jin, X.; Zhuang, F.; Wang, S.; He, Q.; and Shi, Z. 2013. Shared structure learning for multiple tasks with multiple views. In *ECML-PKDD*.

Jin, X.; Zhuang, F.; Xiong, H.; Du, C.; Luo, P.; and He, Q. 2014. Multi-task multi-view learning for heterogeneous tasks. In *CIKM*.

Jin, X.; Zhuang, F.; Pan, S. J.; Du, C.; Luo, P.; and He, Q. 2015. Heterogeneous multi-task semantic feature learning for classification. In *CIKM*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv:1607.01759*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.

Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. *ICLR*.

Lau, J. H.; Newman, D.; and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*.

Levy, O., and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. In *NIPS*.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*.

Liang, D.; Krishnan, R. G.; Hoffman, M. D.; and Jebara, T. 2018. Variational autoencoders for collaborative filtering. In *WWW*.

Liu, B.; Li, Y.; Sun, Z.; Ghosh, S.; and Ng, K. 2018a. Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In *AAAI*.

Liu, Y.; Gu, Z.; Ko, T. H.; and Hua, K. A. 2018b. Learning perceptual embeddings with two related tasks for joint predictions of media interestingness and emotions. In *ICMR*.

Luo, D.; Xu, H.; Zhen, Y.; Ning, X.; Zha, H.; Yang, X.; and Zhang, W. 2015. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *IJCAI*.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *KDD*.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR*.

Mao, C.; Yao, L.; and Luo, Y. 2019. Medgcn: Graph convolutional networks for multiple medical tasks. *arXiv:1904.00326*.

Matveeva, I. 2006. Document representation and multilevel measures of document similarity. In *NAACL*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *ICML*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *EMNLP*.

Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; and Eisenstein, J. 2018. Explainable prediction of medical codes from clinical text. *arXiv:1802.05695*.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *NAACL*.

Ogura, Y., and Kobayashi, I. 2013. Text classification based on the latent topics of important sentences extracted by the pagerank algorithm. In *ACL Workshop*.

Onan, A.; Korukoğlu, S.; and Bulut, H. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*.

Rekabsaz, N.; Mitra, B.; Lupu, M.; and Hanbury, A. 2017. Toward incorporation of relevant documents in word2vec. *arXiv:1707.06598*.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*.

Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; and LeCun, Y. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229*.

Shen, D.; Wang, G.; Wang, W.; Min, M. R.; Su, Q.; Zhang, Y.; Li, C.; Henao, R.; and Carin, L. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *ACL*.

Shi, H.; Xie, P.; Hu, Z.; Zhang, M.; and Xing, E. P. 2017. Towards automated ICD coding using deep learning. *arXiv:1711.04075*.

Srivastava, A., and Sutton, C. 2017. Autoencoding variational inference for topic models. *arXiv:1703.01488*.

Suo, Q.; Ma, F.; Canino, G.; Gao, J.; Zhang, A.; Veltri, P.; and Agostino, G. 2017. A multi-task framework for monitoring health conditions via attention-based recurrent neural networks. In *AMIA annual symposium proceedings*.

Suo, Q.; Ma, F.; Yuan, Y.; Huai, M.; Zhong, W.; Gao, J.; and Zhang, A. 2018. Deep patient similarity learning for personalized healthcare. *IEEE transactions on nanobioscience*.

Thrun, S. 1996. Is learning the n-th thing any easier than learning the first? In *NIPS*.

Wallach, H. M.; Mimno, D. M.; and McCallum, A. 2009. Rethinking lda: Why priors matter. In *NIPS*.

Wang, W.; Gan, Z.; Wang, W.; Shen, D.; Huang, J.; Ping, W.; Satheesh, S.; and Carin, L. 2017. Topic compositional neural language model. *NAACL*.

Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; and Carin, L. 2018a. Joint embedding of words and labels for text classification. *ACL*.

Wang, W.; Pu, Y.; Verma, V. K.; Fan, K.; Zhang, Y.; Chen, C.; Rai, P.; and Carin, L. 2018b. Zero-shot learning via class-conditioned deep generative models. In *AAAI*.

Wang, W.; Tao, C.; Gan, Z.; Wang, G.; Chen, L.; Zhang, X.; Zhang, R.; Yang, Q.; Henao, R.; and Carin, L. 2019. Improving textual network learning with variational homophilic embeddings. *NeurIPS*.

Wang, X.; Wang, F.; and Hu, J. 2014. A multi-task learning framework for joint disease risk prediction and comorbidity discovery. In *ICPR*.

Xu, H.; Wu, W.; Nemati, S.; and Zha, H. 2017. Patient flow prediction via discriminative learning of mutually-correcting processes. *IEEE transactions on Knowledge and Data Engineering*.

Xu, H.; Wang, W.; Liu, W.; and Carin, L. 2018. Distilled wasserstein learning for word embedding and topic modeling. In *NIPS*.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *WWW*.

Yang, X.; Kim, S.; and Xing, E. P. 2009. Heterogeneous multitask learning with joint sparsity constraints. In *NIPS*.

Yao, L.; Mao, C.; and Luo, Y. 2019. Graph convolutional networks for text classification. *AAAI*.

Yu, K.; Tresp, V.; and Schwaighofer, A. 2005. Learning gaussian processes from multiple tasks. In *ICML*.

Zhang, J., and Huan, J. 2012. Inductive multi-task learning with multiple view data. In *KDD*.

Zhang, Y., and Yeung, D.-Y. 2011. Multi-task learning in heterogeneous feature spaces. In *AAAI*.

Zhang, J.; Ghahramani, Z.; and Yang, Y. 2008. Flexible latent variable models for multi-task learning. *Machine Learning*.