Knowledge Distillation via Constrained Variational Inference

Ardavan Saeedi,[1] Yuria Utsumi,[2] Li Sun,[3] Kayhan Batmanghelich,[3] Li-wei H. Lehman[2]

[1] Hyperfine*
[2] Massachusetts Institute of Technology
[3] University of Pittsburgh
av.saeedi@gmail.com, yutsumi@mit.edu, lis118@pitt.edu, kayhan@pitt.edu, lilehman@mit.edu

Abstract

Knowledge distillation has been used to capture the knowledge of a teacher model and distill it into a student model with some desirable characteristics such as being smaller, more efficient, or more generalizable. In this paper, we propose a framework for distilling the knowledge of a powerful discriminative model such as a neural network into commonly used graphical models known to be more interpretable (e.g., topic models, autoregressive Hidden Markov Models). Posterior of latent variables in these graphical models (e.g., topic proportions in topic models) is often used as feature representation for predictive tasks. However, these posterior-derived features are known to have poor predictive performance compared to the features learned via purely discriminative approaches. Our framework constrains variational inference for posterior variables in graphical models with a similarity preserving constraint. This constraint distills the knowledge of the discriminative model into the graphical model by ensuring that input pairs with (dis)similar representation in the teacher model also have (dis)similar representation in the student model. By adding this constraint to the variational inference scheme, we guide the graphical model to be a reasonable density model for the data while having predictive features which are as close as possible to those of a discriminative model. To make our framework applicable to a wide range of graphical models, we build upon the Automatic Differentiation Variational Inference (ADVI), a black-box inference framework for graphical models. We demonstrate the effectiveness of our framework on two real-world tasks of disease subtyping and disease trajectory modeling.

1 Introduction

Distilling knowledge of a teacher model in a student model was originally motivated by compressing larger neural networks into smaller ones (Hinton, Vinyals, and Dean 2015). However, later it has been applied to a diverse set of areas such as adversarial defense (Papernot et al. 2016) or privileged learning (Lopez-Paz et al. 2015). The distinguishing factor among these applications is the desirable characteristic of the student model (e.g. higher inference speed, smaller size). In all these applications, the student model mimics the performance of the teacher model while maintaining the desirable characteristic.

In this paper, we propose a framework for distilling the knowledge of a discriminative model into a probabilistic graphical model. Probabilistic graphical models such as topic models or autoregressive hidden Markov models (AR-HMM) have been widely used for building density models of observed data. Two factors that have helped their widespread adoption are their simplicity and the possibility of bypassing custom inference for them by using probabilistic programming languages. In these models, posterior of the latent variables (e.g. topic proportions in a topic model) or a function of it is often used as low-dimensional feature representation for a downstream task such as predicting labels associated with each data point (Halpern et al. 2012; Lehman et al. 2015b). However, as shown by Halpern et al. (2012) and Hughes et al. (2018), this two-stage process of extracting features and then training a discriminative model, performs subpar compared to purely discriminative approaches. Semi-supervised variants of these models, where labels and observations are modeled jointly, have been developed but their performance is not significantly different from the two-stage approaches (Hughes et al. 2018).

Our goal is to enhance the feature representation of these graphical models so they incorporate the knowledge of their discriminative counterparts while being reasonable density models for the data. Optimizing for matching the observed data can be achieved by maximizing the variational lower bound of the marginal likelihood. To incorporate the knowledge of a powerful discriminative model (i.e. our teacher model), we add a knowledge distillation constraint to the variational inference optimization objective. This constraint ensures that we have the best generative model for the data while having feature space which is close to that of a discriminative model.

Given the distinct nature of the feature space between the teacher and student models in our framework, we propose to use the similarity-preserving knowledge distillation scheme introduced by Tung and Mori (2019) as our constraint. This scheme, instead of the common approach of matching the smoothed class scores of the teacher and student models, matches the pairwise similarity matrix in the student model with that matrix in the teacher model. A pairwise similarity matrix consists of pairwise distances between the feature representations of inputs to a model. As mentioned by Tung and Mori (2019), this scheme is inspired by the observation

*Work is not related to the research done at Hyperfine.

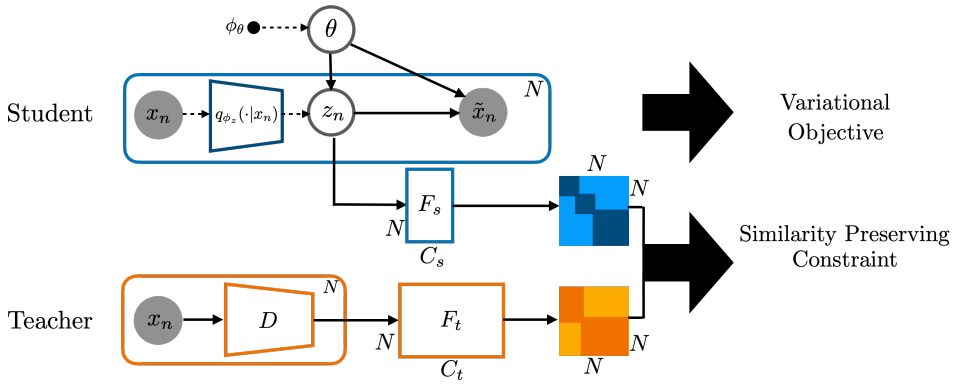


Figure 1: *Knowledge Distillation via Constrained Variational Inference*: Illustration of our framework for distilling knowledge from a teacher model (D) with high predictive power into a probabilistic graphical model with local variables z, global variables θ, and N observations. Dashed lines represent relations in the variational approximation. F_t and F_s are feature representation matrices from the teacher and student models. Pairwise similarity matrices are built based on these matrices for the teacher and student models. To make our approach flexible, we use a black-box variational inference scheme for our student model.

that semantically similar inputs tend to have similar feature representations. Fig. 1 illustrates our framework.

To make our framework applicable to a broader range of probabilistic models, we build it upon the Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al. 2017). ADVI is a black-box variational inference method that only requires defining a probabilistic model and a dataset, and it is adopted in probabilistic programming languages such as Stan (Carpenter et al. 2017). ADVI, in its standard form, marginalizes over local latent variables (e.g. cluster assignments in a Gaussian mixture model); this is not a limitation if we are only interested in global latent variables such as means and variances of the mixture components in Gaussian mixtures. However, since we are interested in using the posterior distribution of a local latent variable (or a function of it) as a feature representation in our method, we need to modify ADVI to support our use case. We utilize the Autoencoding Variational Bayes (AEVB) framework (Kingma and Welling 2013) and inference amortization for this purpose. The combination of ADVI with the AEVB framework gives us the flexibility to support a wide range of probabilistic graphical models.

We demonstrate the flexibility of our method by applying it to two real-world tasks of disease subtyping in Chronic Obstructive Pulmonary Disease (COPD) and disease trajectory modeling in MIMIC-III dataset (Johnson et al. 2016). We show knowledge distillation in probabilistic graphical models can improve their predictive performance while not degrading their generative performance.

2 Related Work

Knowledge Distillation Knowledge distillation has been widely used in the neural networks literature for capturing the knowledge of a teacher model to train a student model which has some desirable characteristics. The desirable characteristic has originally been efficiency (Hinton, Vinyals, and Dean 2015; Ahn et al. 2019; Hegde et al. 2020); however, later more diverse set of applications have emerged. For instance, in privileged learning (Lopez-Paz et al. 2015), a teacher model with access to privileged data is distilled to train an unprivileged student model. Li et al. (2017) utilized knowledge distillation for learning from noisy labels. Recently, Ravina et al. (2021), have proposed an approach for distilling interpretable models into human readable code where the desirable characteristic is human-readability and the target is a concise human-readable code. In this paper, the desirable characteristic is interpretability and our target model is a commonly used graphical model (e.g. topic model or an AR-HMM). Nanfack, Temple, and Frénay (2021) also proposed a framework for distilling the knowledge of a black-box model into an interpretable model; however, in contrast to our model, they used a decision rule-based explanation instead of commonly used graphical models.

Black-Box Variational Inference Black-box variational inference (BBVI) methods generalize variational inference and typically only require computing the gradients of the variational approximation (Ranganath, Gerrish, and Blei 2014; Salimans and Knowles 2014). Kingma and Welling (2013) simplify the optimization process using the reparameterization trick. Kucukelbir et al. (2017) build upon these works and ADVI, a general framework for data analysis which only requires a probabilistic model and a dataset. There are also BBVI methods that are developed for specific domains; for instance, Archer et al. (2015) introduce a BBVI variant for state-space models, or Ambrogioni et al. (2021) propose an approach for BBVI with structured variational family. Our method combines ADVI and AEVB in order to support inference in graphical models with both local and global latent variables.

(Semi-)Supervised Learning with Probabilistic Graphical Models Many models have been developed with the goal of being a reasonable density model of observed data while having high predictive power. The basic approach of using the inferred latent variables from a graphical model as features for training a discriminative model have been

employed in various applications such as predicting psychological state (Resnik, Garron, and Resnik 2013), patient's health in ICU (Lehman et al. 2012, 2015a), or patient monitoring (Lehman, Mark, and Nemati 2018). This approach has limited success due to the fact that the inferred features may not be relevant for the prediction task. Hence, other approaches have been proposed that model the data and labels jointly by including label generation as part of the generative process (Blei and McAuliffe 2010; Li, Ouyang, and Zhou 2015; Chen et al. 2015). Our approach can improve the performance of these methods by distilling the knowledge of a pre-trained discriminative model into the generative model. Hoyle, Goel, and Resnik (2020) also use a knowledge distillation approach for improving the performance of topic models via pretrained transformers. Their approach is specifically designed for topic models, while our approach can be applied to any probabilistic graphical model.

Constrained Inference To improve upon the (semi-)supervised approaches, constrained inference or posterior regularization approaches have been proposed (Hughes et al. 2018; Zhu, Ahmed, and Xing 2012; Zhu, Chen, and Xing 2014). These approaches are highly specific to their applications and constrain the posterior by enforcing explicit performance constraints. In contrast, our approach is general and can even be used in combination with these methods. Furthermore, instead of constraining the posterior by performance, we constrain it by the feature space of another pre-trained discriminative model. In another less related area, constrained variational inference has also been used for encoding human knowledge into the inference procedure (Unhelkar and Shah 2019).

3 Method

The class of student models we support consists of graphical models with (1) local latent variables $Z = z_{1:N}$, (2) global latent variables θ, and (3) a dataset with N observations $X = x_{1:N}$. A local latent variable z_n encodes the hidden structure that governs the n^{th} observation and the global variables θ are the model parameters that are provided with some prior distribution. As shown by Hoffman et al. (2013), these graphical models are general and cover widely used families of models such as sequential models, mixture models or topic models. In Section 4, we provide examples of our framework applied to different families of graphical models.

The graphical model defines a joint likelihood over the observations and the latent variables $p(x, z, \theta)$. Identifying patterns in the data and prediction tasks usually amounts to computing the posterior $p(z, \theta|x)$ in these models. Given the intractability of the posterior for many graphical models, approximate inference techniques such as variational inference have been proposed. However, to avoid custom optimization routines for variational inference, BBVI frameworks have been developed and are commonly used in probabilistic inference software packages such as Stan (Carpenter et al. 2017). In such frameworks, the goal is to maximize the Evidence Lower BOund (ELBO) with respect to the variational parameters $\phi = \{\phi_\theta, \phi_z\}$:

$$\mathcal{L}(\phi_\theta, \phi_z; x) \triangleq \mathbb{E}_{q_\phi}[\log p(x, z, \theta) - \log q_{\phi_\theta}(\theta) - \log q_{\phi_z}(z|x)], \quad (1)$$

without painstaking derivations of the variational update equations. As demonstrated by Kucukelbir et al. (2017), sidestepping these derivations allows these frameworks to be applicable to much larger families of graphical models that do not assume conditional or full conjugacy. To expand the applicability of our approach, we develop our framework based on the ideas from BBVI frameworks. In particular, we utilize ADVI for approximating the posterior of our global variables θ and *recognition* network as posterior approximator for our local variables z. This allows us to avoid a parameter space that grows (at least) linearly with the number of observations for our local variables.

3.1 Global Latent Variables θ

We follow the approach proposed by Kucukelbir et al. (2017) for our global variables which we assume are continuous and hence differentiable. This limitation can be alleviated by adopting some of the latest techniques for gradient estimation in models with discrete latent variables (Tucker et al. 2017; Grathwohl et al. 2018; Kool, van Hoof, and Welling 2020).

ADVI recipe for developing a general variational inference algorithm is to transform the latent variables θ (with K dimensions) such that they live in the real coordinate space $\mathbb{R}^K$: $T : \text{supp}(p(\theta)) \to \mathbb{R}^K$, where T is the transformation and supp is the support of the distribution. This implicitly defines the variational approximation in the original space as $q(\theta; \phi_\theta) = q(T(\theta); \phi_\theta)|\det(J_T(\theta))|$. Consequently, ADVI can choose the variational distribution independent of the generative model.

After the transformation, one can assume a factorized Gaussian distribution as the variational approximation for the transformed latent variables $\Theta = T(\theta)$:

$$q(\Theta; \phi_\theta) = \mathcal{N}(\Theta; \boldsymbol{\mu}, \text{diag}(\exp(\boldsymbol{\omega})^2)), \quad (2)$$

where $\phi_\theta = (\mu_1, \cdots, \mu_K, \omega_1, \cdots, \omega_K)$ are the variational parameters in the unconstrained space of $\mathbb{R}^{2K}$. We will rewrite the variational objective in Eq. 1 with this transformation for the global latent variables in Section 3.3.

3.2 Local Latent Variables z

To have an efficient input-dependent variational approximation, we employ a recognition network for amortizing inference of our local latent variables z (with L dimensions). Our recognition network maps observations x into the approximate posterior $q(z|x)$. However, to ensure the applicability of the recognition network to various types of latent variables, we follow the ADVI recipe for transforming the local latent variables into a variable in real coordinate space: $\zeta = T(z)$. Similar to Eq. 2, we assume a factorized Gaussian distribution for the variational distribution of ζ:

$$q_{\phi_z}(\zeta|x) = \mathcal{N}(\boldsymbol{\mu}_{\phi_z}(x), \text{diag}(\exp(\boldsymbol{\omega}_{\phi_z}(x))^2)), \quad (3)$$

where ϕ_z denotes the parameter set of the recognition network.

3.3 Variational Objective $\mathcal{L}(\phi; x)$

Given the transformations in Sections 3.1 and 3.2, we rewrite the ELBO in real coordinate space for datapoint x_i as follows:

$$\begin{aligned}\mathcal{L}(\phi_\theta, \phi_z; x_i) \triangleq \mathbb{E}_{q_\phi}[&\log p(x_i, T^{-1}(\Theta), T^{-1}(\zeta)) \\ &+ \log|\det(J_{T^{-1}}(\Theta))| \\ &+ \log|\det(J_{T^{-1}}(\zeta))|] \\ &+ \mathbb{H}(q_{\phi_\theta}(\Theta)) + \mathbb{H}(q_{\phi_z}(\zeta|x_i)).\end{aligned}$$

This can be optimized in the real coordinate space by differentiating $\mathcal{L}(\phi_\theta, \phi_z; x_i)$ with respect to ϕ_θ and ϕ_z. To push the gradient operation inside the expectation, we use the "reparameterization trick" (Kingma and Welling 2013) and write the expectation in terms of a standard Gaussian density with $L + K$ dimensions:

$$\begin{aligned}\mathcal{L}(\phi_\theta, \phi_z; x_i) \triangleq \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{I})}[&\log p(x_i, T^{-1}(\Theta_{\boldsymbol{\epsilon}}), T^{-1}(\zeta_{\boldsymbol{\epsilon}})) \\ &+ \log|\det(J_{T^{-1}}(\Theta_{\boldsymbol{\epsilon}}))| + \log|\det(J_{T^{-1}}(\zeta_{\boldsymbol{\epsilon}}))|] \\ &+ \mathbb{H}(q_{\phi_\theta}(\Theta)) + \mathbb{H}(q_{\phi_z}(\zeta|x_i)),\end{aligned} \tag{4}$$

where $\Theta_{\boldsymbol{\epsilon}} = \text{diag}(\exp(\boldsymbol{\omega})) \odot \boldsymbol{\epsilon}_{1:K} + \boldsymbol{\mu}$, and

$$\zeta_{\boldsymbol{\epsilon}} = \text{diag}(\exp(\boldsymbol{\omega}_{\phi_z}(x_i)) \odot \boldsymbol{\epsilon}_{K+1:} + \boldsymbol{\mu}_{\phi_z}(x_i).$$

3.4 Prediction Based on Local Latent Variables

The local latent variables $z_{1:N}$ inferred via a recognition network are often used as features for another predictive task where we predict labels $y_{1:N}$. Examples of these features include posterior topic proportions in a topic model or marginal posterior of latent states in a hidden Markov model. As mentioned in Section 2, a two-stage process or supervised variants of latent variable models are two possible approaches for prediction based on the latent variables in these models. Hughes et al. (2018) show that constraining the space of a generative model with a prediction-based constraint significantly improves the predictive performance of these models. Our knowledge distillation method, which adds distillation as a constraint to our optimization problem (Eq. 4), is inspired by this observation.

3.5 Knowledge Distillation Constraint

The distinct nature of the representation space between the teacher model–typically a neural network–and the student model calls for a different approach than simply mimicking the teacher model's representation by the student model. Inspired by the observation that semantically similar inputs should generate similar feature representation in a trained neural network, and similar to the approach introduced by Tung and Mori (2019), we propose a similarity-preserving knowledge distillation method. We distill knowledge in the student model such that input pairs that produce (dis)similar feature representations in the teacher model have (dis)similar representations in the student model.

For a dataset of size N, we denote the feature representation extracted from the teacher and the student models by $F^t \in \mathbb{R}^{N \times C_t}$ and $F^s \in \mathbb{R}^{N \times C_s}$, correspondingly. Here C_t and C_s are the sizes of the feature vectors for the teacher and student models. For the student model, we assume each row n of F^s is a function of the corresponding inferred local latent variable: $F^s_{[n,:]} = f(q_{\phi_z}(z_n|x_n))$. Note that this can also be an identity function; for instance, in the case of a topic model we use topic proportions as feature representation in the student model. Following Tung and Mori (2019), we define the knowledge distillation constraint to ensure the differences between the ℓ^2-normalized outer products of F^s and F^t are less than some predefined tolerance level η.

Concretely, we first compute the similarity between the feature representations via a dot product and obtain the following $N \times N$ matrices:

$$\tilde{F^s} = F^s \cdot F^{s\top} \text{ and } \tilde{F^t} = F^t \cdot F^{t\top}. \tag{5}$$

Next, we normalize $\tilde{F^s}$ and $\tilde{F^t}$ by applying ℓ^2 normalization to each row and obtain $\bar{F^s}$ and $\bar{F^t}$. Finally, we define the constraint as: $\frac{1}{N^2}\left\|\bar{F^s} - \bar{F^t}\right\|^2_{\mathbf{F}} < \eta$. Where $\|\cdot\|_{\mathbf{F}}$ denotes the Frobenius norm. We write our constrained optimization problem as:

$$\begin{aligned}&\min_{\phi_\theta, \phi_z} -\mathcal{L}(\phi_\theta, \phi_z; x_i) \\ &\text{s.t. } \frac{1}{N^2}\left\|\bar{F^s} - \bar{F^t}\right\|^2_{\mathbf{F}} < \eta.\end{aligned} \tag{6}$$

Using the Lagrange multipliers, the constrained version of Eq. 1 can be written in an unconstrained format with a multiplier $\gamma_\eta > 0$ corresponding to a tolerance level η:

$$\min_{\phi_\theta, \phi_z} -\mathcal{L}(\phi_\theta, \phi_z; x_i) + \gamma_\eta \frac{1}{N^2}\left\|\bar{F^s} - \bar{F^t}\right\|^2. \tag{7}$$

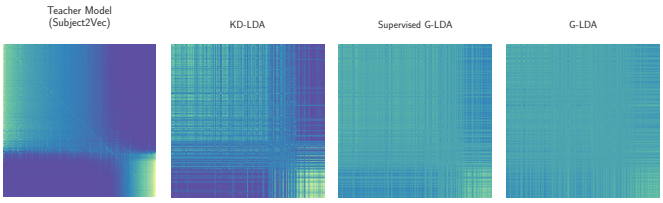
In practice, we use stochastic optimization and subsample data in mini-batches of size B. We estimate the gradients of the objective function for the full dataset based on the mini-batches. Furthermore, since there is no analytic form for the relationship between γ_η and η, we treat γ_η as a hyperparameter.

4 Experiments

4.1 Disease Subtyping in Chronic Obstructive Pulmonary Disease (COPD)

Task COPD, one of the leading causes of death worldwide (World Health Organization 2018), is characterized by inflammation of the airway, and is a highly heterogenous disease (Castaldi et al. 2017; Chen, Xu, and Xiao 2013). Computed tomography (CT) imaging, is used for qualitative and quantitative evaluation of tissue inflammation and destruction in COPD.

Given that there are differences between risk factors of COPD subtypes, understanding disease subtypes is important (Shapiro 2000). Different disease subtypes can manifest themselves as different tissue patterns in a CT image and multiple subtypes may be present simultaneously in a patient (Batmanghelich et al. 2015). Hence, we can view the CT image of a patient as a mixture of typical imaging patterns that are common across the population and apply a topic model with appropriate observation model. Our goal is to identify tissue subtypes which are relevant for predicting clinical measurements indicative of disease severity. The clinical



(a) (b) (c) (d)

Figure 2: *Pairwise similarity matrices for the test dataset of the disease subtyping experiment:* Each row and each column corresponds to a patient in the test dataset. Brighter colors indicate higher similarity values. The distillation constraint encourages the pairwise similarity matrix of the student model (KD-LDA) to be similar to that of the teacher model (Section 3.5). Compared to the G-LDA model (d) which is unsupervised, the matrix for the supervised G-LDA (c) is more similar to that of the teacher.

measurements that we are interested in are four positive real variables, including percent predicted values of Forced Expiratory Volume in one second (`FEV1pp`) and its ratio with Forced vital capacity ($\texttt{FEV}_1/\texttt{FVC}$), percent predicted values of Forced vital capacity (`FVCpp`) and Distance Walked. These clinical-relevant variables are highly correlated with disease severity and are described in details in the appendix. In what follows, we use topic and subtype interchangeably.

Dataset We base our evaluation on a large-scale dataset from COPDGene study (Regan et al. 2011) with lung CT images for 7,292 subjects. To extract features for each subject, we first segment the lung volume into spatially homogeneous regions using the SLIC superpixel segmentation algorithm (Holzer and Donner 2014). Then for each supervoxel, we extract a 32-bin intensity histogram features (Sorensen et al. 2012) and concatenate it with a rotationally invariant descriptor (sHOG) proposed by Liu et al. (2014). This feature representation has been shown to be relevant in characterizing emphysema (Shaker et al. 2010). The details of our data preprocessing is provided in the appendix. We randomly split the data into a 70% train, 15% validation and 15% test splits.

Student model We use a variant of Latent Dirichlet Allocation (LDA) with D-dimensional Gaussian observations as our student model. In other words, each subtype k is represented by a mean $\mu_k \in \mathbb{R}^D$ and a covariance $\Sigma_k \in \mathbb{R}^D \times \mathbb{R}^D$. For a population of S patients with K possible subtypes the n^{th} supervoxel of patient s, v_{sn}, is generated by the following generative process:

$$\begin{aligned}
\pi_s &\sim \text{Dir}(\alpha_0) && s \in \{1 \cdots S\} \\
\mu_k, \Sigma_k &\sim \text{NIW}(\xi_0) && k \in \{1 \cdots K\} \\
z_{sn} &\sim \text{Cat}(\pi_s) && s \in \{1 \cdots S\} \\
v_{sn} &\sim \mathcal{N}(\mu_{z_{sn}}, \Sigma_{z_{sn}}) && s \in \{1 \cdots S\} \quad n \in \{1 \cdots N_s\}
\end{aligned}$$

where $\text{Dir}(\alpha_0)$ is the Dirichlet distribution with concentration parameter α_0, $\text{NIW}(\xi)$ is the Normal-Inverse-Wishart distribution with hyper-parameter ξ, Cat indicates the categorical distribution, π_s is the topic distribution for patient s, z_{sn} represents the topic assignment for supervoxel n of patient s, and N_s is the number of supervoxels in patient s.

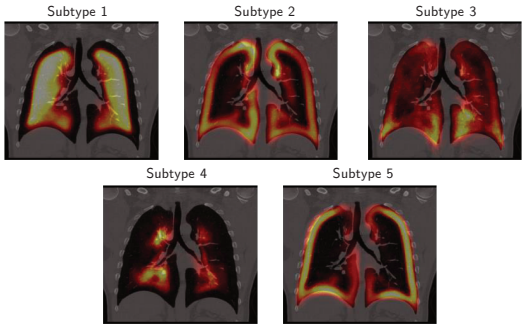
For inference, we use ADVI scheme for global latent variables μ_k and Σ_k, a recognition network for local latent variables π_s, and marginalize over supervoxel's topic assignment z_{sn}. As mentioned in Section 3.1, in our scheme we need a transformation to real coordinate space for the variables not in this space. To transform the π_s variables drawn from the Dirichlet distribution, we apply an inverse stick-breaking transformation which maps the variables on a simplex of K dimensions to an unconstrained space of dimension $K-1$ (Linderman, Johnson, and Adams 2015).

We use the approximate posterior of subtype proportions for patient s (π_s) as feature representation of that patient. Our goal is to predict the clinical measurements mentioned above from π_s for each patient. Our knowledge distillation constraint is also applied to this local variable.

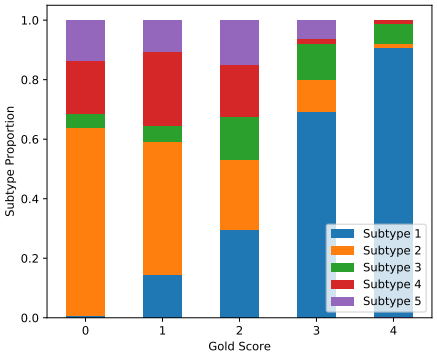
Teacher model For the teacher model we use the *Subject2Vec* (Singla et al. 2018) which is among the best discriminative approaches for predicting disease severity on the COPDGene dataset. The model is inspired by deep sets (Zaheer et al. 2017) and transforms the input set of supervoxels to a fixed-length representation. To aggregate the supervoxels, it adaptively weights each one based on its contribution to the prediction of disease severity. We use the 128-dimensional learned feature vectors from this model for our knowledge distillation constraint. The predictive performance of this model is shown in Table 1 to present the performance upper bound. The pairwise similarity matrices for the student and teacher models along with those for the baselines are presented in Fig. 2.

Baselines We denote the basic model without any knowledge distillation or supervision by *G-LDA*, the model with joint modeling of labels and observations by *Supervised G-LDA*, and the model with knowledge distillation by *KD-LDA*. For predicting the clinical measurements in the G-LDA baseline, we train a linear regression model on the posterior of topic proportions $q(\pi_s|v_{s.})$. For the supervised G-LDA, we model label y_s as $y_s|\pi_s \sim \mathcal{N}(f(\pi_s), \sigma_0)$, where f is a learnable linear function and σ_0 is a hyperparameter. For the KD-LDA, similar to G-LDA, we train a linear model on the inferred topic proportions. See appendix for the hyperparameter setting and details of all the experiments.

Results To evaluate our model we need to show improvement of the predictive performance while not significantly affecting the generative aspect of the model. Table 1, shows that our model outperforms the baselines in terms of the coefficient of determination R^2 of the prediction. To show our method does not significantly affect the generative aspect of the model, we report ELBO in Table 2. While the supervised G-LDA can result in worse ELBO values for some clinical measurements, our method does not have a significant impact on ELBO. We visualize the learned subtypes in Fig. 3a, which shows that different subtypes focus on different anatomical regions. Furthermore, we show the average distributions of subtypes and their relation with disease severity in Fig. 3b. To categorize the disease severity, we use Global Initiative



(a)



(b)

Figure 3: *Disease Subtyping Experiment*: (a) Visualization of spatial average of the learned subtypes across the population shown on a coronal slice of a lung atlas. (b) Subtype proportions averaged over subsets of the population with GOLD score values 0, 1, 2, 3, and 4. Higher values mean more severe disease. All bars have equal sizes but the proportion of subtypes varies. The proportion of subtypes varies. The proportion of subtype 1 increases and subtype 2 decreases as we move from GOLD score 0 to 4 (indicating severely diseased).

	R^2			
Subject-Level Descriptor	FEV1pp	FEV_1/FVC	FVCpp	Distance Walked
Subject2Vec (Teacher)	0.65	0.70	0.28	0.16
G-LDA	0.16 ±0.04	0.30 ±0.05	0.03 ±0.01	0.05 ±0.01
Supervised G-LDA	0.29 ±0.05	0.49 ±0.05	-0.47 ±0.07	0.06 ±0.02
KD-LDA	**0.49 ±0.01**	**0.61 ±0.02**	**0.15 ±0.01**	**0.14 ±0.01**

Table 1: Performance of predicting clinical-relevant variables (FEV1pp , FEV_1/FVC, FVCpp, and distance walked) compared across G-LDA, supervised G-LDA, and our method KD-LDA. Our method outperforms the rest in all clinical metrics. Results are averaged across 5 runs for each method. The teacher (Subject2Vec) model's performance is added as a reference. Note that for G-LDA, we need to train a separate model for each clinical variable.

	ELBO ($\times 10^3$)				
Subject-Level Descriptor	Unsupervised	FEV1pp	FEV_1/FVC	FVCpp	Distance Walked
G-LDA	-2.88 ±0.08	-	-	-	-
Supervised G-LDA	-	-2.72 ±0.01	-2.73 ±0.01	-3.01 ±0.01	-3.09 ±0.01
KD-LDA	-2.73 ±0.01	-	-	-	-

Table 2: ELBO values for the G-LDA, Supervised G-LDA, and KD-LDA. Results are averaged across 5 runs for each method. Note that only the supervised G-LDA has different ELBO values depending on the clinical measurement we want to predict. This is due to the fact that we are modeling the observations and labels jointly in this model. The difference between KD-LDA and G-LDA is not significant.

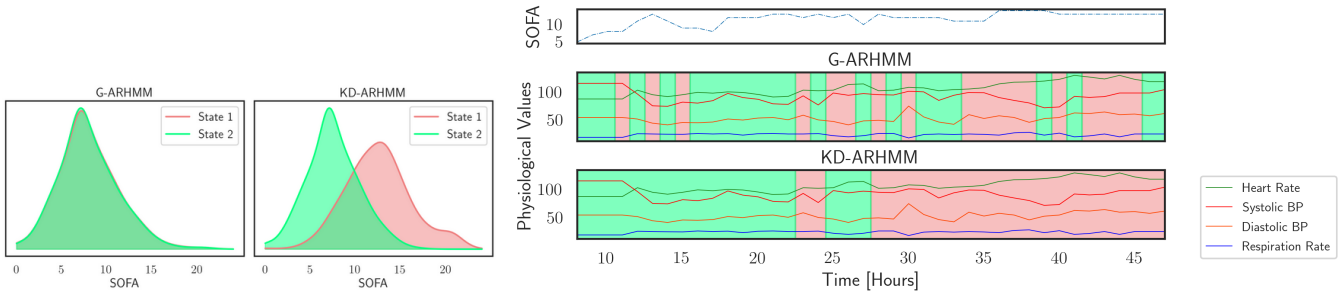
for Obstructive Lung Disease (GOLD) which is a discrete variable between zero and four. Zero is used for people at risk (normal spirometry but chronic symptoms), and 1-4 denote mild to very severe COPD. In Fig. 3b, each bar represents a sub-population of patients with a particular GOLD score and colors within the bar are the average proportion of a subtype within that sub-population. The results in Fig. 3b show that the proportion of subtype 1 increases as we move from GOLD score 0 to 4 (indicating severely diseased). Subtype 2 and 5, in contrast, decrease with increased severity.

4.2 Dynamics Modeling of Patients' Clinical States for Sepsis Monitoring

Task We focus on the task of sepsis disease progression monitoring in the intensive care unit (ICU). Our aim is to identify patients' clinical states of health to generate early alerts of impending physiological deterioration of patients at high-risk of in-hospital mortality. We evaluate the clinical utility of the state marginals learned from our approach in estimating patients' mortality risks. We also demonstrate that the latent states learned through our approach contain clinically rich information that are more informative of patients' end-organ status, as indicated by the widely-used Sequential Organ Failure Assessment (SOFA) score (Vincent et al. 1996), in comparison to the baselines. In what follows, we use latent state and clinical state interchangeably.

Dataset We extract the cohort from MIMIC-III (Johnson et al. 2016), a public, de-identified critical care database. We use the criteria defined by Singer and et al. (2016) for our sepsis cohort and limit to patients with at least 48-hours of ICU data, giving a total cohort of 11648 patients. The dataset includes a total of 29 time-varying physiological and clinical variables. We randomly split the dataset into 70% train, 15% validation, and 15% test splits.

Student model We use an autoregressive hidden Markov Model (AR-HMM) as our student model. For a population of S patients, each with covariates of dimension D and length T_s, we model the cohort times series as an order 1 switching vector autoregressive process with K possible latent states, with the k-th state parameterized by $\theta_k = \{A_k, b_k, \Sigma_k\}$, where AR coefficients $A_k \in \mathbb{R}^D \times \mathbb{R}^D$, bias vector $b_k \in \mathbb{R}^D$, noise covariance $\Sigma_k \in \mathbb{R}^D \times \mathbb{R}^D$. Let x_t^s represents the covariate of patient s at time step t. The generative process is as follows:



(a) (b)

Figure 4: *Dynamics Modeling of Clinical States Experiment:* (a) Distribution of SOFA scores, a clinical measure of patients' end-organ function, for learned latent states. Higher SOFA indicates worse end-organ function. The distributions are more distinguishable for KD-ARHMM than for G-ARHMM, which indicates our proposed approach infers clinically meaningful latent states. (b) Segmentation of inferred clinical states for a high-risk patient with worsening end-organ function. KD-ARHMM segmentations correlate more with SOFA score trajectory.

$$\begin{aligned} \pi_k &\sim \text{Dir}(\alpha) && k \in \{1 \cdots K\} \\ z_t^s &\sim \text{Cat}(\pi_{z_{t-1}^s}) \\ x_t^s &\sim \mathcal{N}(A_{z_t^s} x_{t-1}^s + b_{z_t^s}, \Sigma_{z_t^s}), \end{aligned}$$

where π_k is the state-specific transition distribution for state k, and z_t^s represents the latent state for covariates of patient s at time t. For a patient s, we use the average marginal of approximate posterior distribution for the clinical states (z^s) as feature representation of that patient. Our goal is to model the progression of patients' health states, and predict in-hospital mortality from z^s for each patient.

Teacher model For the teacher model, we use a long short-term memory network (LSTM) which has shown to be effective in prediction-based healthcare related tasks (Tomašev et al. 2019; Xiao, Choi, and Sun 2018; Choi et al. 2016). We use the hidden state representations from this model at the final timestep for our knowledge distillation constraint. The predictive performance of this model is shown in Table 3 to present the performance upper bound.

Baselines We denote the basic model without any knowledge distillation or supervision by *G-ARHMM*, the model with joint modeling of labels and observations by *Supervised G-ARHMM*, and the model with knowledge distillation by *KD-ARHMM*. For predicting in-hospital mortality in the G-ARHMM and KD-ARHMM, we follow a two-stage process similar approach to the one in Section 4.1. For the supervised G-ARHMM, we have the same generative process for the labels as supervised G-LDA.

Results Table 3 shows that our model outperforms the baselines in terms of AUROC of the prediction, with 95% confidence intervals (DeLong, DeLong, and Clarke-Pearson 1988). In terms of ELBO our method even improves the generative performance. In Fig. 4a and Fig. 4b we show that KD-ARHMM infers clinically meaningful latent states in comparison to the baseline. See appendix for the pairwise similarity matrices for the student and teacher models and the hyperparameter settings.

	AUROC (95% CI)	ELBO ($\times 10^5$)
LSTM (Teacher)	0.71 (0.68, 0.74)	N/A
G-ARHMM	0.56 (0.52, 0.59)	-55.24
Supervised G-ARHMM	0.56 (0.52, 0.59)	-55.24
KD-ARHMM	**0.65 (0.61, 0.68)**	**-3.94**

Table 3: Performance of predicting in-hospital mortality compared across G-ARHMM, supervised G-ARHMM, and our method KD-ARHMM. The teacher (LSTM) model's performance is added as a reference.

5 Concluding Remarks

We introduced a framework for knowledge distillation in probabilistic graphical models by adding a similarity-preserving constraint to the variational objective function. The constraint encourages the pairwise similarity matrix of the student model (i.e. graphical model) to be similar to that of the teacher model (i.e. a discriminative model with superior predictive performance). To make the framework general, we employed BBVI framework and combined ADVI and AEVB to handle both local and global variables. We demonstrated the performance of our model compared to reasonable baselines and showed that improvement in the predictive performance in our model does not significantly impact the generative aspect of it. Following black-box variational inference means we are vulnerable to its known weaknesses: underestimating the posterior variance, sensitivity to initialization, and amortization gap (Cremer, Li, and Duvenaud 2018). On the other hand, building our method based on BBVI means we can benefit from the new developments that try to tackle these issues (e.g. (Giordano, Broderick, and Jordan 2018; Cremer, Li, and Duvenaud 2018)). An avenue for future research and further theoretical analysis could be understanding these limitations in the context of knowledge distillation.

Acknowledgements

We thank Adam Dejl for his contribution. LL was in part funded by NIH grants R01-EB030362, R01-EB017205 and MIT-IBM Watson AI Lab.

References

Ahn, S.; Hu, S. X.; Damianou, A.; Lawrence, N. D.; and Dai, Z. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9163–9171.

Ambrogioni, L.; Lin, K.; Fertig, E.; Vikram, S.; Hinne, M.; Moore, D.; and Gerven, M. 2021. Automatic structured variational inference. In *International Conference on Artificial Intelligence and Statistics*, 676–684. PMLR.

Archer, E.; Park, I. M.; Buesing, L.; Cunningham, J.; and Paninski, L. 2015. Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*.

Batmanghelich, N. K.; Saeedi, A.; Cho, M.; Estepar, R. S. J.; and Golland, P. 2015. Generative method to discover genetically driven image biomarkers. In *International Conference on Information Processing in Medical Imaging*, 30–42. Springer.

Blei, D. M.; and McAuliffe, J. D. 2010. Supervised topic models. *arXiv preprint arXiv:1003.0783*.

Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M. A.; Guo, J.; Li, P.; and Riddell, A. 2017. Stan: a probabilistic programming language. *Grantee Submission*, 76(1): 1–32.

Castaldi, P. J.; Benet, M.; Petersen, H.; Rafaels, N.; Finigan, J.; Paoletti, M.; Marike Boezen, H.; and et al. 2017. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax*, 72(11): 998–1006.

Chen, J.; He, J.; Shen, Y.; Xiao, L.; He, X.; Gao, J.; Song, X.; and Deng, L. 2015. End-to-end learning of LDA by mirror-descent back propagation over a deep architecture. *arXiv preprint arXiv:1508.03398*.

Chen, X.; Xu, X.; and Xiao, F. 2013. Heterogeneity of chronic obstructive pulmonary disease: from phenotype to genotype. *Frontiers of medicine*, 7(4): 425–32.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29.

Cremer, C.; Li, X.; and Duvenaud, D. 2018. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, 1078–1086. PMLR.

DeLong, E.; DeLong, D.; and Clarke-Pearson, D. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44: 837–845.

Giordano, R.; Broderick, T.; and Jordan, M. I. 2018. Covariances, robustness and variational Bayes. *Journal of machine learning research*, 19(51).

Grathwohl, W.; Choi, D.; Wu, Y.; Roeder, G.; and Duvenaud, D. 2018. Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*.

Halpern, Y.; Horng, S.; Nathanson, L. A.; Shapiro, N. I.; and Sontag, D. 2012. A comparison of dimensionality reduction techniques for unstructured clinical text. In *Icml 2012 workshop on clinical data analysis*, volume 6.

Hegde, S.; Prasad, R.; Hebbalaguppe, R.; and Kumar, V. 2020. Variational student: Learning compact and sparser networks in knowledge distillation framework. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3247–3251. IEEE.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5).

Holzer, M.; and Donner, R. 2014. Over-Segmentation of 3D Medical Image Volumes based on Monogenic Cues. *Cvww*, (JANUARY 2014): 35–42.

Hoyle, A. M.; Goel, P.; and Resnik, P. 2020. Improving Neural Topic Models using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hughes, M. C.; Hope, G.; Weiner, L.; McCoy Jr, T. H.; Perlis, R. H.; Sudderth, E. B.; and Doshi-Velez, F. 2018. Semi-Supervised Prediction-Constrained Topic Models. In *AISTATS*, 1067–1076.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L. H.; 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*, 3(1): 1–9.

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

Kool, W.; van Hoof, H.; and Welling, M. 2020. Estimating gradients for discrete random variables by sampling without replacement. *arXiv preprint arXiv:2002.06043*.

Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; and Blei, D. M. 2017. Automatic Differentiation Variational Inference. *J. Mach. Learn. Res.*, 18(1): 430–474.

Lehman, L.; Johnson, M.; Nemati, S.; Adams, R.; and Mark, R. 2015a. Bayesian nonparametric learning of switching dynamics in cohort physiological time series: application in critical care patient monitoring. *Advanced State Space Methods for Neural and Clinical Data*, 257.

Lehman, L. H.; Adams, R. P.; Mayaud, L.; Moody, G. B.; Malhotra, A.; Mark, R. G.; and Nemati, S. 2015b. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE journal of biomedical and health informatics*, 19(3): 1068–1076.

Lehman, L. H.; Mark, R. G.; and Nemati, S. 2018. A Model-Based Machine Learning Approach to Probing Autonomic Regulation from Nonstationary Vital-Sign Time Series. *IEEE journal of biomedical and health informatics*, 22(1): 56–66.

Lehman, L. H.; Saeed, M.; Long, W.; Lee, J.; and Mark, R. G. 2012. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In *Proceedings of the AMIA Annual Symposium*, 505–511.

Li, X.; Ouyang, J.; and Zhou, X. 2015. Supervised topic models for multi-label classification. *Neurocomputing*, 149: 811–819.

Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; and Li, L.-J. 2017. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1910–1918.

Linderman, S. W.; Johnson, M. J.; and Adams, R. P. 2015. Dependent Multinomial Models Made Easy: Stick Breaking with the PóLya-gamma Augmentation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 3456–3464.

Liu, K.; Skibbe, H.; Schmidt, T.; Blein, T.; Palme, K.; Brox, T.; and Ronneberger, O. 2014. Rotation-Invariant HOG Descriptors Using Fourier Analysis in Polar and Spherical Coordinates. *International Journal of Computer Vision*, 106(3): 342–364.

Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; and Vapnik, V. 2015. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*.

Nanfack, G.; Temple, P.; and Frénay, B. 2021. Global explanations with decision rules: a co-learning approach. In *Uncertainty in Artificial Intelligence*, 589–599. PMLR.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, 582–597. IEEE.

Ranganath, R.; Gerrish, S.; and Blei, D. 2014. Black box variational inference. In *Artificial intelligence and statistics*, 814–822. PMLR.

Ravina, W.; Sterling, E.; Oryeshko, O.; Bell, N.; Zhuang, H.; Wang, X.; Wu, Y.; and Grushetsky, A. 2021. Distilling Interpretable Models into Human-Readable Code. *arXiv preprint arXiv:2101.08393*.

Regan, E. A.; Hokanson, J. E.; Murphy, J. R.; Make, B.; Lynch, D. A.; Beaty, T. H.; Curran-Everett, D.; Silverman, E. K.; and Crapo, J. D. 2011. Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1): 32–43.

Resnik, P.; Garron, A.; and Resnik, R. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1348–1353.

Salimans, T.; and Knowles, D. 2014. On using control variates with stochastic approximation for variational Bayes. *arXiv preprint arXiv*, 1401.

Shaker, S. B.; Bruijne, M. D.; Sorensen, L.; Shaker, S. B.; and De Bruijne, M. 2010. Quantitative analysis of pulmonary emphysema using local binary patterns. *Medical Imaging, IEEE Transactions on*, 29(2): 559–569.

Shapiro, S. D. 2000. Evolving concepts in the pathogenesis of chronic obstructive pulmonary disease. *Clin Chest Med*, 21(4): 621–632.

Singer, M.; and et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8): 801–810.

Singla, S.; Gong, M.; Ravanbakhsh, S.; Sciurba, F.; Poczos, B.; and Batmanghelich, K. N. 2018. Subject2Vec: generative-discriminative approach from a set of image patches to a vector. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 502–510. Springer.

Sorensen, L.; Nielsen, M.; Lo, P.; Ashraf, H.; Pedersen, J. H.; and De Bruijne, M. 2012. Texture-based analysis of COPD: A data-driven approach. *IEEE Transactions on Medical Imaging*, 31(1): 70–78.

Tomašev, N.; Glorot, X.; Rae, J. W.; Zielinski, M.; Askham, H.; Saraiva, A.; Mottram, A.; Meyer, C.; Ravuri, S.; Protsyuk, I.; et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767): 116–119.

Tucker, G.; Mnih, A.; Maddison, C. J.; Lawson, D.; and Sohl-Dickstein, J. 2017. REBAR: Low-variance, unbiased gradient estimates for discrete variable models.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1365–1374.

Unhelkar, V. V.; and Shah, J. A. 2019. Learning models of sequential decision-making with partial specification of agent behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2522–2530.

Vincent, J.; Moreno, R.; Takala, J.; Willatts, S.; De Mendonça, A.; Bruining, H.; Reinhart, C.; Suter, P.; and Thijs, L. 1996. The SOFA (Sepsis-Related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22: 707–710.

World Health Organization. 2018. The top 10 causes of death. https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Online; accessed 12-June-2018].

Xiao, C.; Choi, E.; and Sun, J. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10): 1419–1428.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R.; and Smola, A. 2017. Deep sets. *arXiv:1703.06114*.

Zhu, J.; Ahmed, A.; and Xing, E. P. 2012. MedLDA: maximum margin supervised topic models. *the Journal of machine Learning research*, 13(1): 2237–2278.

Zhu, J.; Chen, N.; and Xing, E. P. 2014. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1): 1799–1847.