Clinical Risk Prediction with Temporal Probabilistic Asymmetric Multi-Task Learning

A. Tuan Nguyen [*] [†] [1, 4], **Hyewon Jeong** [*] [1], **Eunho Yang** [1, 2, 3], **Sung Ju Hwang** [1, 2, 3] [‡]

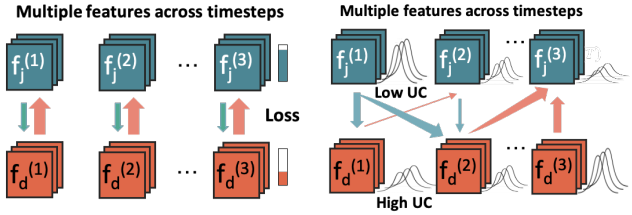
[1] School of Computing, Korea Advanced Institute of Science and Technology,
[2] AI Graduate School, Korea Advanced Institute of Science and Technology
[3] Aitrics,
[4] Department of Computer Science, University of Oxford
tuan.nguyen@cs.ox.ac.uk, {jhw162, eunhoy, sjhwang82}@kaist.ac.kr

Abstract

Although recent multi-task learning methods have shown to be effective in improving the generalization of deep neural networks, they should be used with caution for safety-critical applications, such as clinical risk prediction. This is because even if they achieve improved task-average performance, they may still yield degraded performance on individual tasks, which may be critical (e.g., prediction of mortality risk). Existing asymmetric multi-task learning methods tackle this *negative transfer* problem by performing knowledge transfer from tasks with low loss to tasks with high loss. However, using loss as a measure of reliability is risky since low loss could result from overfitting. In the case of time-series prediction tasks, knowledge learned for one task (e.g., predicting the sepsis onset) at a specific timestep may be useful for learning another task (e.g., prediction of mortality) at a later timestep, but lack of loss at each timestep makes it challenging to measure the reliability at each timestep. To capture such dynamically changing asymmetric relationships between tasks in time-series data, we propose a novel temporal asymmetric multi-task learning model that performs knowledge transfer from certain tasks/timesteps to relevant uncertain tasks, based on the feature-level uncertainty. We validate our model on multiple clinical risk prediction tasks against various deep learning models for time-series prediction, which our model significantly outperforms without any sign of negative transfer. Further qualitative analysis of learned knowledge graphs by clinicians shows that they are helpful in analyzing the predictions of the model.

Introduction

Multi-task learning (MTL) (Caruana 1997) is a method to train a model, or multiple models jointly for multiple tasks to obtain improved generalization, by sharing knowledge among them. One of the most critical problems in multi-task learning is the problem known as *negative transfer*, where unreliable knowledge from other tasks adversely affects the target task. This negative transfer could be fatal for safety-critical applications such as clinical risk prediction, where we cannot risk losing performance on any of the tasks. To prevent negative transfer, researchers have sought ways to allow knowledge transfer only among closely related tasks, by either identifying the task groups or learning optimal sharing structures among tasks (Duong et al. 2015; Misra et al. 2016). However, it is not only the task relatedness that matters, but also the relative reliability of the task-specific knowledge. Recent asymmetric multi-task learning (AMTL) models (Lee, Yang, and Hwang 2016, 2017) tackle this challenge by allowing tasks with low loss to transfer more.

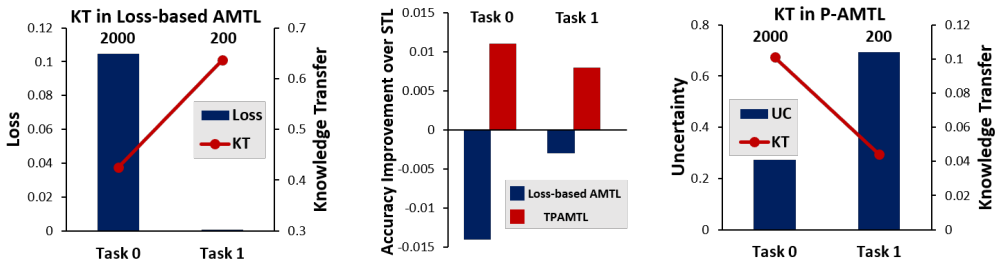


(a) Loss-based AMTL (b) TPAMTL

Figure 1: Temporal Probabilistic Asymmetric Multi-Task Learning. **(a)** Existing AMTL models (Lee, Yang, and Hwang 2016, 2017) utilize task loss to perform static knowledge transfer (KT) from one task to another; thus it cannot capture dynamically changing relationships between timesteps and tasks in the time-series domain. **(b)** Our model performs dynamic KT among tasks and across timesteps based on the feature-level uncertainty (UC).

While the asymmetric knowledge transfer between tasks is useful, it does not fully exploit the asymmetry in time-series analysis, which has an additional dimension of the time axis. With time-series data, knowledge transfer direction may need to be different depending on the timestep. For instance, suppose that we predict an event of infection or mortality within 48 hours of admission in intensive care units (ICU) based on electronic health records (EHR). At earlier timesteps, prediction of *Infection* may be more reli-

[*] Equal contribution
[†] This work was done while the author is in KAIST
[‡] Correspondence to: A. Tuan Nguyen, Hyewon Jeong, and Sung Ju Hwang.



(a) (b) (c)

Figure 2: Experiment with a small, imbalanced number of training instances (Task 0: 2000 instances, and Task1: 200 instances) explains the failure case of (a) loss-based AMTL which fails to reliably transfer knowledge between tasks, compared to (c) probabilistic asymmetric multi-task learning. Also please refer to Table 1 for full experiment.

able than *Mortality*, where we may want knowledge transfer to happen from task *Infection* to *Mortality*; at later timesteps, we may want the next situation to happen. Moreover, knowledge transfer may happen across timesteps. For example, a high risk of *Infection* in early timestep will alert high risk of *Mortality* at later timesteps. To exploit such temporal relationships between tasks, we need a model that does not perform static knowledge transfer between two tasks (Figure 1a), but transfers knowledge across the timesteps of two different tasks, while dynamically changing the knowledge transfer amount and direction at each timestep (Figure 1b). To this end, we propose a multi-task learning framework for time-series data, where each task not only learns its own latent features at each timestep but also leverages aggregated latent features from the other tasks at the same or different timesteps via attention allocation (Figure 3).

Yet this brings in another challenge. On what basis should we promote asymmetric knowledge transfer? For asymmetric knowledge transfer between tasks, we could use task loss as a proxy of knowledge reliability (Lee, Yang, and Hwang 2016, 2017). However, loss is not a direct measure of reliability, as loss might not be available at every step for time-series prediction. Also, a model trained with few instances (Task 1 in Figure 2a - 2c) may have a small loss and thus transfer more knowledge to other tasks (Figure 2a), but the knowledge from this model could be highly biased and unreliable as it may have overfitted (Figure 2b). Thus, we propose a novel probabilistic Bayesian framework for asymmetric knowledge transfer, which leverages feature-level *uncertainty*, instead of task loss, to measure the reliability of the knowledge (Figure 1b). Basically, if a latent feature learned at a certain timestep has large uncertainty, our model will allocate small attention values for the feature; that is, the attention will be attenuated based on the uncertainty, where knowledge transfers from the task with low uncertainty to high uncertainty (Figure 2c; Please see Table 1).

We experimentally validate our *Temporal Probabilistic Asymmetric Multi-Task Learning (TP-AMTL)* model on **four clinical risk prediction datasets** against multiple baselines. The results show that our model obtains significant improvements over strong multi-task learning baselines with no **negative transfer** on any of the tasks (Table 2). We further show that both the asymmetric knowledge transfer between tasks at two different timesteps and the uncertainty-based attenuation of attention weights are effective in improving generalization. Finally, with the actual knowledge transfer graph plotted with uncertainty obtained for each timestep, we could interpret the model behaviors according to actual clinical events in clinical risk prediction tasks (Figure 4, Supplemenatary Figure 13, 14). This interpretability makes it more suitable for clinical risk prediction in real-world situations.

Our contribution is threefold:

- We propose a **novel probabilistic formulation for asymmetric knowledge transfer**, where the amount of knowledge transfer depends on the feature level uncertainty.
- We extend the framework to **an asymmetric multi-task learning framework for time-series analysis**, which utilizes feature-level uncertainty to perform knowledge transfer among tasks and across timesteps, thereby exploiting both the task-relatedness and temporal dependencies.
- We validate our model on **four clinical risk prediction tasks** against ten baselines, which it significantly outperforms with no **negative transfer**. With the help of clinicians, we further analyze the learned knowledge transfer graph to discover meaningful relationships between clinical events.

Related Work

Multi-Task Learning

While the literature on multi-task learning (Caruana 1997; Argyriou, Evgeniou, and Pontil 2008) is vast, we selectively mention the prior works that are closely related to ours. Historically, multi-task learning models have focused on *what to share* (Yang and Hospedales 2016a,b; Ruder et al. 2017), as the jointly learned models could share instances, parameters, or features (Kang, Grauman, and Sha 2011; Kumar and Daume III 2012; Maurer, Pontil, and Romera-Paredes 2013). With deep learning, multi-task learning can be implemented

rather straightforwardly by making multiple tasks to share the same deep network. However, since solving different tasks will require diversified knowledge, complete sharing of the underlying network may be suboptimal and brings in a problem known as *negative transfer*, where certain tasks are negatively affected by knowledge sharing. To prevent this, researchers are exploring more effective knowledge sharing structures. Soft parameter sharing method with regularizer (Duong et al. 2015) can enforce the network parameters for each task to be similar, while a method to learn the optimal combination of shared and task-specific representations is also proposed (Misra et al. 2016) in computer vision. Losses can be weighed based on the uncertainty of the task in a multi-task framework (Kendall, Gal, and Cipolla 2018), reducing negative transfer from uncertain tasks. While finding a good sharing structure can alleviate negative transfer, negative transfer will persist if we perform symmetric knowledge transfer among tasks. To resolve this symmetry issue, the asymmetric MTL model with inter-task knowledge transfer (Lee, Yang, and Hwang 2016) was proposed, which allows task-specific parameters for tasks with smaller loss to transfer more. Lee, Yang, and Hwang (2017) proposed a model for asymmetric task-to-feature transfer that allows reconstructing features with task-specific features while considering their loss, which is more suitable for deep neural networks and scalable.

Clinical Time-Series Analysis

While our method is generic and applicable to any time-series prediction task, we mainly focus on clinical time-series analysis in this paper. Multiple clinical benchmark datasets (Citi and Barbieri 2012; Johnson et al. 2016) have been released and publicly available. Also, several recent works have proposed clinical prediction benchmarks with publicly available datasets (Che et al. 2018; Harutyunyan et al. 2019; Johnson, Pollard, and Mark 2017; Pirracchio 2016; Purushotham et al. 2017). We construct our datasets and tasks specific to our problem set (Experiments section), in part referring to previous benchmark tasks. Furthermore, there has been some progress on this topic recently, mostly focusing on the interpretability and reliability of the model. An attention-based model (Choi et al. 2016) that generates attention for both the timesteps (hospital visits) and features (medical examination results) was proposed to provide interpretations of the predictions. However, attentions are often unreliable since they are learned in a weakly-supervised manner, and a probabilistic attention mechanism (Heo et al. 2018) was also proposed to obtain reliable interpretation and prediction that considers uncertainty as to how to trust the input. Our work shares the motivation with these prior works as we target interpretability and reliability. Recently, SAnD (Song et al. 2018)) proposes to utilize a self-attention architecture for clinical prediction tasks, and Adacare (Ma et al. 2020) proposes scale-adaptive recalibration module with dilated convolution to capture time-series biomarker features. Yet, they are inherently susceptible to negative transfer as all tasks share a single base network (Table 2).

Approach

Probabilistic Asymmetric Multi-Task Learning

In this section, we describe our framework of probabilistic asymmetric multi-task learning (P-AMTL) in a general setting. Suppose that we have D tasks with datasets $\{\mathbf{X}_d, \mathbf{Y}_d\}_{d=1}^D$, in which the sets $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_D$ can be identical, overlapping or even disjoint. We further suppose that we have D different probabilistic networks $\{p_d(.)\}_{d=1}^D$, each of which generates high-level latent features of task d (task-specific) via $\mathbf{Z}_d \sim p_d(\mathbf{X}_d)$. In a single-task learning setting, these latent features $\mathbf{Z}_d$ are in turn used to make predictions for task d. However, in our asymmetric multi-task learning framework, we want to borrow some learned features from other tasks to share knowledge and to improve generalization performance. Specifically, in order to perform prediction for task d, we leverage latent features learned from other tasks, $\mathbf{Z}_{j,d} \sim p_j(\mathbf{X}_d), \forall j \neq d$. Given the source features $\mathbf{Z}_{j,d}$ and the target features $\mathbf{Z}_d$, the model needs to decide on the following:

1) The amount of knowledge to transfer Existing asymmetric multi-task learning models (Lee, Yang, and Hwang 2016, 2017) often use task loss to decide on the amount of knowledge transfer, in a way that tasks with low training loss are allowed to transfer more, while tasks with high loss only receive knowledge from other tasks. However, the task loss may not be a proper measure of the knowledge from the task and also unavailable in some cases (Figure 2a, cases in Introduction). To overcome these limitations, we propose to learn the amount of knowledge transfer based on the feature-level UC. Our model learns the transfer weight from $\mathbf{Z}_{j,d}$ to $\mathbf{Z}_d$ by a small network $F_{j,d}$ (Equation 1). This learnable network takes both $\mathbf{Z}_{j,d}$, $\mathbf{Z}_d$ and their variance $\boldsymbol{\sigma}_{j,d}^2$ and $\boldsymbol{\sigma}_d^2$ as its input. Note that if the variance is not available from the output of $\{p_d(.)\}_{d=1}^D$ directly, we can perform Monte-Carlo sampling k times on $\mathbf{Z}_{j,d}$ and $\mathbf{Z}_d$ to compute the estimates of variances. In practice, we can implement each $F_{j,d}$ as a multi-layer perceptron with the input as the concatenation of $\mathbf{Z}_{j,d}, \mathbf{Z}_d, \boldsymbol{\sigma}_{j,d}^2$ and $\boldsymbol{\sigma}_d^2$.

$$\alpha_{j,d} = F_{j,d}(\mathbf{Z}_{j,d}, \mathbf{Z}_d, \boldsymbol{\sigma}_{j,d}^2, \boldsymbol{\sigma}_d^2) \tag{1}$$

2) The form of transferred knowledge Since the learned features for different tasks may have completely different representations, directly adding $\alpha_{j,d}\mathbf{Z}_{j,d}$ to $\mathbf{Z}_d$ would be sub-optimal. For this combining process, we train two additional networks G_k^1 and G_k^2 for each task k where G_k^1 is used to convert the learned task-specific features from $p_k(.)$ to a shared latent space and G_k^2 is used to convert the features from that shared latent space back to the task-specific latent space. Finally, we compute the combined feature map for task d as (Figure 3 (Right)):

$$\mathbf{C}_d = \mathbf{Z}_d + G_d^2\left(\sum_{j \neq d} \alpha_{j,d} * G_j^1(\mathbf{Z}_{j,d})\right) \tag{2}$$

The combined feature map $\mathbf{C}_d$ can then be used for the final prediction for task d. The combined feature maps for all other tasks are computed in the same manner.

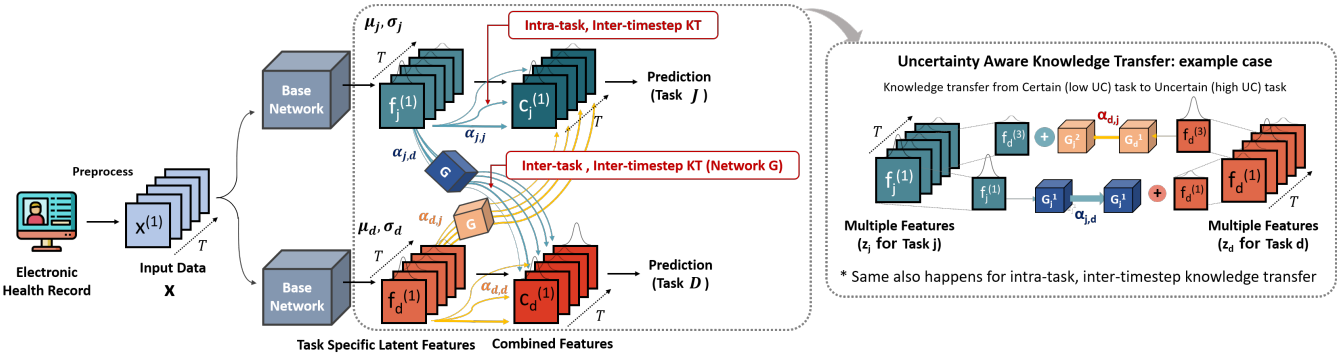


Figure 3: Temporal probabilistic asymmetric knowledge transfer. This figure illustrates how we apply the probabilistic asymmetric knowledge transfer between tasks at the same timestep and across different timesteps. (Right) Features of task j at timestep 1 is more reliable than features of task d at timestep 1, so the model will learn to transfer more from task j to task d and transfer less from task d to task j.

Temporal Probabilistic Asymmetric Multi-Task Learning

We now apply our probabilistic asymmetric multi-task learning framework for the task of time-series prediction. Our goal is to jointly train time-series prediction models for multiple tasks at once. Suppose that we are given training data for D tasks, $\mathbb{D} = \{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_D, \mathbf{Y}_D)\}$. Further suppose that each data instance $\mathbf{x}$ where $\mathbf{x} \in \mathbf{X}_d$ for some task d, consists of T timesteps. That is, $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(T)})$, where $\mathbf{x}^{(t)} \in \mathbb{R}^m$ denotes the data instance for the timestep t. Here we assume the number of timesteps T is identical across all tasks for simplicity, but there is no such restriction in our model. Additionally, y_d is the label for task d; $y_d \in \{0, 1\}$ for binary classification tasks, and $y_d \in \mathbb{R}$ for regression tasks. Given time-series data and tasks, we want to learn the task-specific latent features for each task and timestep, and then perform asymmetric knowledge transfer between them. Our framework is comprised of the following components:

Shared Low-Level Layers We allow our model to share low-level layers for all tasks in order to learn a common data representation before learning task-specific features. At the lowest layer, we have a shared linear data embedding layer to embed the data instance for each timestep into a continuous shared feature space. Given a time-series data instance $\mathbf{x}$, we first linearly transform the data point for each timestep t, $\mathbf{x}^{(t)} \in \mathbb{R}^m$, which contains m variables:

$$(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, ..., \mathbf{v}^{(T)}) = \mathbf{v} = \mathbf{x}\mathbf{W}_{emb} \in \mathbb{R}^{T \times k} \quad (3)$$

where $\mathbf{W}_{emb} \in \mathbb{R}^{m \times k}$ and k is the hidden dimension. This embedded input is then fed into shared RNN layer for pre-processing:

$$\mathbf{r} = (\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, ..., \mathbf{r}^{(T)}) = RNN(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, ..., \mathbf{v}^{(T)}) \quad (4)$$

Task- and Timestep Embedding Layers ('Base Network' in Figure 3) After embedding and pre-processing the input into a continuous space, we further encode them into task- and timestep-specific features. Since hard-sharing layers may result in negative transfer between tasks, we allow *separate* embedding layers for each task to encode task-specific knowledge in our 'Base Network'. For each task d, the separate network consists of L feed-forward layers, which learn disentangled knowledge for each timestep. These L feed-forward layers for task embedding can be formulated as:

$$\mathbf{h}_d = \sigma((...\sigma(\sigma(\mathbf{r}\mathbf{W}_d^1 + \mathbf{b}_d^1)\mathbf{W}_d^2 + \mathbf{b}_d^2)...)\mathbf{W}_d^L + \mathbf{b}_d^L) \in \mathbb{R}^{T \times k} \quad (5)$$

where $\mathbf{W}_d^i \in \mathbb{R}^{k \times k}, \mathbf{b}_d^i \in \mathbb{R}^k$ and σ is a non-linear activation function (e.g. leaky relu).

Modeling feature-level uncertainty While the embedding above can capture knowledge for each task and timestep, we want to further model their uncertainties as well, to measure the reliability of the knowledge captured. Towards this objective, we model the latent variables as probabilistic random variables, with two types of UC (Kendall and Gal 2017): 1) *epistemic uncertainty*, which comes from the model's unreliability from the lack of training data, and 2) *aleatoric uncertainty*, that comes from the inherent ambiguity in the data. We capture the former by using dropout variational inference (Gal and Ghahramani 2016), and the latter by explicitly learning the model variance as a function of the input (Figure 3). Suppose that we have the task-specific latent features: $\mathbf{z}_d \sim p(\mathbf{z}_d|\mathbf{x}, \boldsymbol{\omega})$ where $\boldsymbol{\omega}$ is the set of all parameters. This can be formulated as below:

$$\mathbf{z}_d|\mathbf{x}, \boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{z}_d; \boldsymbol{\mu}_d, diag(\boldsymbol{\sigma}_d^2)\right) \quad (6)$$

$$\boldsymbol{\mu}_d = \sigma(\mathbf{h}_d\mathbf{W}_d^{\mu} + \mathbf{b}_d^{\mu}) \quad (7)$$

$$\boldsymbol{\sigma}_d = softplus(\mathbf{h}_d\mathbf{W}_d^{\sigma} + \mathbf{b}_d^{\sigma}) \quad (8)$$

As mentioned before, we use dropout approximation (Gal and Ghahramani 2016) with parameter M as the variational distribution $q_M(\boldsymbol{\omega})$ to approximate $p(\boldsymbol{\omega}|\mathbb{D})$.

Asymmetric knowledge transfer across tasks and time steps Now we apply the proposed probabilistic asymmet-

ric knowledge transfer method to perform knowledge transfer across timesteps, both within each task and across tasks, to exploit intra- and inter-task temporal dependencies. In order to transfer knowledge from task j to task d with temporal dependencies, we allow the latent features of task d at time step t ($\mathbf{f}_d^{(t)}$, with $\mathbf{z}_d = (\mathbf{f}_d^{(1)}, \mathbf{f}_d^{(2)}, ..., \mathbf{f}_d^{(T)})$) to obtain knowledge from task j at all previous time steps (Figure 3), and then combine them into a single feature map $\mathbf{C}_d^{(1)}, \mathbf{C}_d^{(2)}, ..., \mathbf{C}_d^{(T)}$:

$$\mathbf{C}_d^{(t)} = \mathbf{f}_d^{(t)} + G_d^2 \left(\sum_{j=1}^{D} \sum_{i=1}^{t} \alpha_{j,d}^{(i,t)} * G_j^1 \left(\mathbf{f}_j^{(i)} \right) \right) \forall t \quad (9)$$

As mentioned in the previous subsection, the transfer weight $\alpha_{j,d}^{(i,t)}$ is computed by a network $F_{j,d}$ with input $\mathbf{f}_d^{(t)}$, $\mathbf{f}_j^{(i)}$ and their variance $\boldsymbol{\sigma}_d^{(t)}$, $\boldsymbol{\sigma}_j^{(i)}$ (again, we perform MC sampling to get the variance).

Here, we choose to constrain the knowledge transfer to happen only from past to future timesteps because of the time complexity at inference time. With our proposed model, for **each update** at the clinical environment in an online manner, we only need to transfer the knowledge from previous time steps to the current one, making the complexity to be $\mathbf{O}(\mathbf{T})$. This is on a par with other models like RETAIN (Choi et al. 2016) or UA (Heo et al. 2018), making it highly scalable (Table 1 of **Supplementary File**). However, if we allow the knowledge to transfer from future timestep to past timestep, we also need to update the knowledge at previous timesteps for a single update. The time complexity of the model in this case is $O(T^2)$, which is undesirable. In the ablation study, we show that this constraint also brings in a small performance gain. The total complexity of the **whole** training or inference is still $O(T^2)$ due to the inter-timestep transfer, but this is on par with state-of-the-art models such as Transformer (Vaswani et al. 2017) and SAnD (Song et al. 2018).

Finally, we use the combined features $\mathbf{C}_d^{(1)}, \mathbf{C}_d^{(2)}, ..., \mathbf{C}_d^{(T)}$, which contain temporal dependencies among tasks, for prediction for each task d. We use an attention mechanism

$$\boldsymbol{\beta}_d^{(t)} = tanh \left(\mathbf{C}_d^{(t)} \mathbf{W}_d^{\beta} + \mathbf{b}_d^{\beta} \right) \quad \forall t \in \{1, 2, ..., T\} \quad (10)$$

where $\mathbf{W}_d^{\beta} \in \mathbb{R}^{k \times k}$ and $\mathbf{b}_d^{\beta} \in \mathbb{R}^k$. Then the model can perform prediction as

$$p_d = p(\widehat{y}_d | \mathbf{x}) = Sigmoid \left(\frac{1}{T} \left(\sum_{t=1}^{T} \boldsymbol{\beta}_d^{(t)} \odot \mathbf{v}^{(t)} \right) \mathbf{W}_d^o + b_d^o \right) \quad (11)$$

for classification tasks, where $\odot$ denotes the element-wise multiplication between attention ${\beta_d}^{(t)}$ and shared input embedding $\mathbf{v}^{(t)}$ (from Eq. 3), $\mathbf{W}_d^o \in \mathbb{R}^{k \times 1}$ and $b_d^o \in \mathbb{R}^1$. Predictions for other tasks are done similarly. Note that our model does not require each instance to have the labels for every tasks. We can simply maximize the likelihood $p(y_d|\mathbf{x})$ whenever the label y_d is available for input x for task d. Furthermore, our model does not require the instances to have the same number of timesteps T.

The loss function of our objective function is:

$$\sum_{x,y} \left[- \sum_{d \in Ta(x)} (y_d \log p_d + (1 - y_d) \log(1 - p_d)) \right] + \beta_{w\_decay} ||\theta||_2^2 \quad (12)$$

where $Ta(x)$ is the set of tasks which we have the labels available for data instance x, β_{w_decay} is the coefficient for weight decay and θ is the whole parameters of our model.

Experiments [1]

Probabilistic Asymmetric Multi-Task Learning

We first validate the effectiveness of the uncertainty-based knowledge transfer for asymmetric multi-task learning, using a non-temporal version of our model. We use a variant (UdeM 2014) of the MNIST dataset (LeCun and Cortes 2010) which contains images of handwritten digits 0-9 with random rotations and background noise. From this dataset, we construct 5 tasks; each task is a binary classification of determining whether the given image belongs to class 0-4. We sample 5000, 5000, 1000, 1000, and 500 examples for task 0, 1, 2, 3, and 4 respectively, such that asymmetric knowledge transfer becomes essential to obtain good performance on all tasks. As for the base network, we use a multi-layer perceptron, which outputs mean and variance of the task-specific latent features.

1) Single Task Learning (STL) learns an independent model for each task.
2) Multi Task Learning (MTL) learns a single base network with 5 disjoint output layers for the 5 tasks.
3) AMTL-Loss. A model that is similar to P-AMTL, with the transfer weight from task j to task d learned by a network $F_{j,d}$ with the average task loss over all instances as the input.
4) P-AMTL. Our probabilistic asymmetric MTL model.

Results (Table 1) show that MTL outperforms STL, but suffers from negative transfer (Task 4, underlined in Table 1). AMTL-Loss underperforms MTL, which shows that the loss is not a good measure of reliability; a model that is overfitted to a task will have a small loss, but its knowledge may be unreliable (Figure 2a). Finally, our model outperforms all baselines without any sign of negative transfer, demonstrating the superiority of uncertainty-based knowledge transfer.

Clinical Risk Prediction from EHR

Clinical risks can be defined in various ways, but in this paper, we define *clinical risks* as the existence of the event (e.g., Heart Failure, Respiratory Failure, Infection, Mortality) that may lead to deterioration of patients' health condition within a given time window (e.g., 48 hour).

[1] For the details on our experimental settings and additional results on two datasets (**MIMIC III - Heart Failure** and **Respiratory Failure**), please see the **supplementary file**.

Models	Task 0	Task 1	Task 2	Task 3	Task 4	Average
STL	0.7513±0.02	**0.7253±0.01**	**0.5401±0.01**	0.5352±0.02	0.6639±0.01	0.6432±0.01
MTL	0.8266±0.01	0.7021±0.01	**0.5352±0.01**	**0.5987±0.01**	0.6203±0.02	0.6565±0.01
AMTL-Loss	0.7317±0.02	**0.7236±0.01**	0.5309±0.01	0.5166±0.02	0.6698±0.01	0.6345±0.01
P-AMTL (ours)	**0.8469±0.01**	**0.7267±0.01**	**0.5382±0.01**	**0.5950±0.01**	**0.6822±0.01**	**0.6778±0.01**

Table 1: AUROC for the MNIST-Variation Experiment

Tasks and Datasets We experiment on four datasets where we compile for clinical risk prediction from two open-source EHR datasets. Every dataset used in this paper contain tasks with clear temporal dependencies between them. The input features have been pre-selected by the clinicians, since the excessive features were not clinically meaningful, and may harm the prediction performance. Missing feature values are imputed with zero values. Pleare refer to the **supplementary file** for the explanation and evaluation on two datasets **MIMIC III - Heart Failure** and **Respiratory Failure**.

1) MIMIC III - Infection. We compile a dataset out of the MIMIC III dataset (Johnson et al. 2016), which contains records of patients admitted to ICU of a hospital. We use records of patients over the age 15, where we hourly sample to construct 48 timesteps from the first 48 hours of admission. Following clinician's guidelines, we select 12 infection-related variables for the features at each timestep (See Table 11, 12 of **supplementary file**). Tasks considered for this dataset are the clinical events before and after infection; *Fever* (Task 1) as the sign of infection with elevated body temperature, *Infection* (Task 2) as the confirmation of infection by the result of microbiology tests, and finally, *Mortality* (Task 3) as a possible outcome of infection (See Figure 1 of **supplementary file** for the task configuration). After pre-processing, approximately 2000 data points with a sufficient amount of features were selected, which was randomly split to approximately $1000/500/500$ for training/validation/test.

2) PhysioNet (Citi and Barbieri 2012). Total of $4,000$ ICU admission records were included, each containing 48 hours of records (sampled hourly) and 31 physiological signs including variables displayed in Table 4. We select 29 infection-related variables for the features available at each timestep (See **supplementary file**). Task used in the experiment includes four binary classification tasks, namely, 1) *Stay*<3: whether the patient would stay in ICU for less than three days, 2) *Cardiac:* whether the patient is recovering from cardiac surgery, 3) *Recovery:* whether the patient is staying in Surgical ICU to recover from surgery, and 4) *Mortality prediction (Mortality)* (See Figure 2 of **supplementary file** for the task configuration). We use a random split of $2800/400/800$ for training/validation/test.

Baselines We compared the single- and multi- task learning baselines to see the effect of negative transfer. Please see the **supplementary file** for descriptions of the baselines, experimental details, and the hyper-parameters used.

Single Task Learning (STL) Baselines
1) STL-LSTM. The single-task learning method with long short-term memory network to capture temporal dependencies.
2) STL-Transformer. Similar STL setting with **1)**, but with Transformer (Vaswani et al. 2017) as the base network.
3) STL-RETAIN (Choi et al. 2016). The attentional RNN for interpretability of clinical prediction with EHR.
4) STL-UA (Heo et al. 2018). Uncertainty-Aware probabilistic attention model.
5) STL-SAnD (Choi et al. 2016). Self-attention model for multi-task time series prediction.
6) STL-AdaCare (Ma et al. 2020) Feature-adaptive modeling with squeeze and excitation block, based on dilated convolution network.

Multi Task Learning (MTL) Baselines.
MTL baselines are the naive hard-sharing multi-task learning method where all tasks share the same network except for the separate output layers for prediction.
7) MTL-LSTM, 8) MTL-Transformer, 9) MTL-RETAIN, 10) MTL-UA, 11) MTL-SAnD, 12) AdaCare Multi-task learning setting with 7) LSTM, 8) Transformer (Vaswani et al. 2017), 9) RETAIN (Choi et al. 2016), 10) UA (Heo et al. 2018), 11) SAnD (Song et al. 2018), 12) AdaCare (Ma et al. 2020) as the base network, respectively.
13) AMTL-LSTM (Lee, Yang, and Hwang 2016). This learns the knowledge transfer graph between task-specific parameters shared across all timesteps with static knowledge transfer between tasks based on the task loss (Figre 1a).
14) MTL-RETAIN-Kendall (Kendall, Gal, and Cipolla 2018). The uncertainty-based loss-weighing scheme with base MTL-RETAIN.
15) TP-AMTL. Our probabilistic temporal AMTL model that performs both intra- and inter-task knowledge transfer.

Quantitative Evaluation We first evaluate the prediction accuracy of the baseline STL and MTL models and ours on the four clinical time-series datasets, by measuring the Area Under the ROC curve (AUROC) (MIMIC-III Infection and PhysioNet (Table 2)). We observe that hard-sharing MTL models outperform STL on some tasks, but suffers from performance degeneration on others (underlined numbers in Table 2 and 3), which shows a clear sign of negative transfer. MTL models especially work poorly on MIMIC-III infection, which has clear temporal relationships between tasks. Probabilistic models (e.g., **UA** (Heo et al. 2018)) generally outperform their deterministic counterparts (e.g., **RETAIN** (Choi et al. 2016)). However, **MTL-RETAIN-Kendall** (Kendall, Gal, and Cipolla 2018), which learns the weight for each task loss based on uncertainty, significantly

	Models	Fever	Infection	Mortality	Average
STL	LSTM	0.6738 ± 0.02	0.6860 ± 0.02	0.6373 ± 0.02	0.6657 ± 0.02
	Transformer	$\mathbf{0.7110 \pm 0.01}$	0.6500 ± 0.01	0.6766 ± 0.01	0.6792 ± 0.01
	RETAIN	0.6826 ± 0.01	0.6655 ± 0.01	0.6054 ± 0.02	0.6511 ± 0.01
	UA	0.6987 ± 0.02	0.6504 ± 0.02	0.6168 ± 0.05	0.6553 ± 0.02
	SAnD	0.6958 ± 0.02	0.6829 ± 0.01	0.7073 ± 0.02	0.6953 ± 0.01
	AdaCare	0.6354 ± 0.02	0.6256 ± 0.03	0.6217 ± 0.01	0.6275 ± 0.08
MTL	LSTM	0.7006 ± 0.03	0.6686 ± 0.02	0.6261 ± 0.03	0.6651 ± 0.02
	Transformer	0.7025 ± 0.01	0.6479 ± 0.02	$\underline{0.6420 \pm 0.02}$	0.6641 ± 0.02
	RETAIN	0.7059 ± 0.02	0.6635 ± 0.01	0.6198 ± 0.05	0.6630 ± 0.02
	UA	$\mathbf{0.7124 \pm 0.01}$	0.6489 ± 0.02	0.6325 ± 0.04	0.6646 ± 0.02
	SAnD	0.7041 ± 0.01	0.6818 ± 0.02	0.6880 ± 0.01	0.6913 ± 0.01
	AdaCare	$\underline{0.5996 \pm 0.01}$	0.6163 ± 0.02	0.6283 ± 0.01	0.6148 ± 0.00
	AMTL-LSTM	0.6858 ± 0.01	0.6773 ± 0.01	0.6765 ± 0.01	0.6798 ± 0.01
	RETAIN-Kendall	0.6938 ± 0.01	$\underline{0.6182 \pm 0.03}$	0.5974 ± 0.02	0.6364 ± 0.02
	TP-AMTL (our model)	$\mathbf{0.7081 \pm 0.01}$	$\mathbf{0.7173 \pm 0.01}$	$\mathbf{0.7112 \pm 0.01}$	$\mathbf{0.7102 \pm 0.01}$

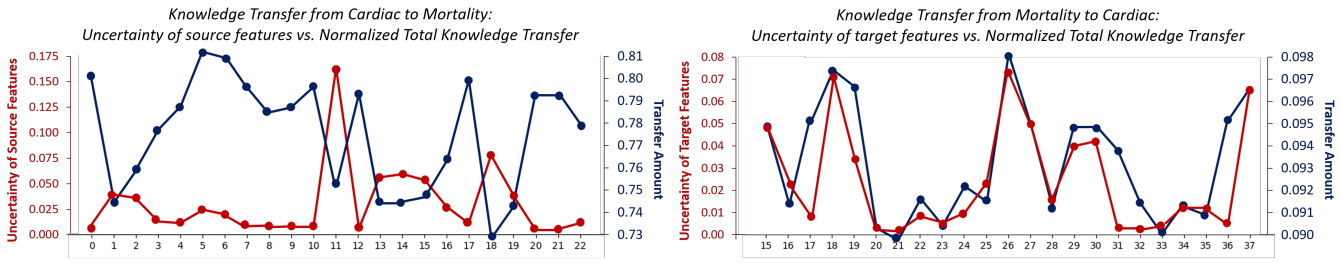
Table 2: Task performance on the MIMIC-III Infection and dataset. We report average AUROC and standard error over five runs (MTL model accuracies lower than those of their STL counterparts are underlined).

	Models	Stay < 3	Cardiac	Recovery	Mortality	Average
STL	LSTM	0.7673 ± 0.09	0.9293 ± 0.01	0.8587 ± 0.01	0.7100 ± 0.01	0.8163 ± 0.03
	Transformer	$\mathbf{0.8953 \pm 0.01}$	0.9283 ± 0.02	0.8721 ± 0.01	0.6796 ± 0.02	0.8380 ± 0.01
	RETAIN	0.7407 ± 0.04	0.9236 ± 0.01	0.8148 ± 0.04	0.7080 ± 0.02	0.7968 ± 0.03
	UA	0.8556 ± 0.02	0.9335 ± 0.01	0.8712 ± 0.01	0.7283 ± 0.01	0.8471 ± 0.01
	SAnD	$\mathbf{0.8965 \pm 0.02}$	0.9369 ± 0.01	0.8838 ± 0.01	0.7330 ± 0.01	0.8626 ± 0.01
	AdaCare	0.7508 ± 0.06	0.8610 ± 0.01	0.7700 ± 0.03	0.6595 ± 0.02	0.7603 ± 0.07
MTL	LSTM	0.7418 ± 0.09	0.9233 ± 0.01	0.8472 ± 0.02	0.7228 ± 0.01	$\underline{0.8088 \pm 0.03}$
	Transformer	$\underline{0.8532 \pm 0.03}$	0.9291 ± 0.01	0.8770 ± 0.01	0.7358 ± 0.01	0.8488 ± 0.01
	RETAIN	0.7613 ± 0.03	$\underline{0.9064 \pm 0.01}$	0.8160 ± 0.04	0.6944 ± 0.03	0.7945 ± 0.03
	UA	0.8573 ± 0.03	0.9348 ± 0.01	0.8860 ± 0.01	$\mathbf{0.7569 \pm 0.02}$	0.8587 ± 0.02
	SAnD	0.8800 ± 0.03	$\mathbf{0.9410 \pm 0.00}$	$\underline{0.8607 \pm 0.01}$	$\mathbf{0.7612 \pm 0.02}$	0.8607 ± 0.06
	AdaCare	0.8746 ± 0.01	$\underline{0.7211 \pm 0.01}$	$\underline{0.6348 \pm 0.02}$	0.7457 ± 0.03	$\underline{0.7440 \pm 0.08}$
	AMTL-LSTM	0.7600 ± 0.08	0.9254 ± 0.01	0.8066 ± 0.01	0.7167 ± 0.01	0.8022 ± 0.03
	RETAIN-Kendall	0.7418 ± 0.02	0.9219 ± 0.02	0.7883 ± 0.03	$\underline{0.6787 \pm 0.02}$	$\underline{0.7827 \pm 0.02}$
	TP-AMTL (our model)	$\mathbf{0.8953 \pm 0.01}$	$\mathbf{0.9416 \pm 0.01}$	$\mathbf{0.9016 \pm 0.01}$	$\mathbf{0.7586 \pm 0.01}$	$\mathbf{0.8743 \pm 0.01}$

Table 3: Task performance on the PhysioNet dataset. We report average AUROC and standard error over five runs (MTL model accuracies lower than those of their STL counterparts are underlined).

underperforms even the **STL-LSTM**, which may be due to the fact that losses in our settings are at almost similar scale unlike with the task losses in (Kendall, Gal, and Cipolla 2018) that have largely different scales. Although the self-attention based model **SAnD** (Song et al. 2018) shows impressive performance on some of the tasks from PhysioNet, it also suffers from performance degeneration in the MTL setting, resulting in lower overall performance. **AMTL-LSTM** (Lee, Yang, and Hwang 2016) improves on some tasks, but degenerates the performance on the others, which we attribute to the fact that it does not consider inter-timestep transfer. Additionally, **AdaCare** with dilated convolution showed severely degenerated performance except for one task. On the other hand, our model, **TP-AMTL**, obtains significant improvements over all STL and MTL baselines on both datasets. It also does not show performance degeneration on any of the tasks, suggesting that it has successfully dealt away with negative transfer in multi-task learning with time-series prediction models. Experimental results on **ablation study** (regarding inter-, intra-task, future-to-past knowledge transfer, various uncertainty types) and additional two datasets (**MIMIC III - Heart Failure** and **Respiratory Failure**) are available in the **supplementary file**, which further supports our model and shows that our model also generalize well to various, larger datasets.

To further analyze the relationships between uncertainty and knowledge transfer, we visualize knowledge transfer from multiple sources (Figure 4a) normalized over the number of targets, and to multiple targets (Figure 4b) normalized over the number of sources, along with their uncertainties. Specifically, the uncertainty of a task at a certain timestep is represented by the average of the variance of all feature distributions. The normalized amount of knowledge transfer from task j at time step t to task d is computed as $(\alpha_{j,d}^{(t,t)}+\alpha_{j,d}^{(t,t+1)}+...+\alpha_{j,d}^{(t,T)})/(T-t+1)$. Similarly, the normalized amount of knowledge transfer to task d at time step t from task j is $(\alpha_{j,d}^{(1,t)}+\alpha_{j,d}^{(2,t)}+...+\alpha_{j,d}^{(t,t)})/t$. We observe that source features with low uncertainty transfer knowledge more, while at the target, features with high uncertainty receive more knowledge transfer. However, note that they



(a) Outgoing Transfer from different Sources

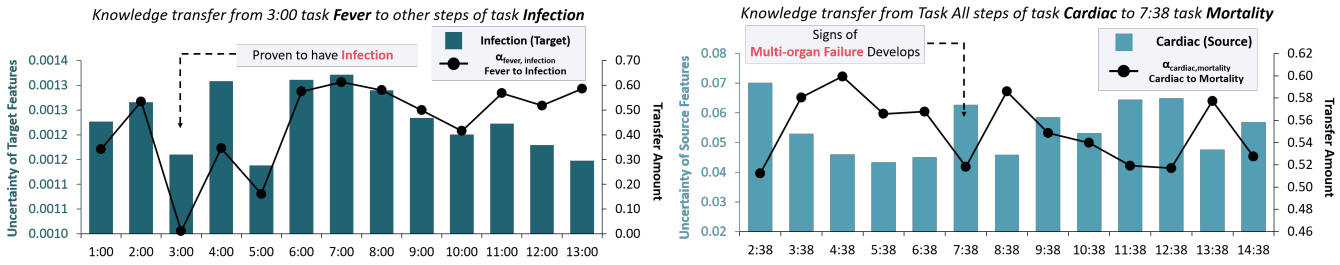
(b) Incoming Transfer to different Targets

Figure 4: Examples showing the relationship between the amount of KT and UC of source and target features. (a) The sources with low UC transfer more knowledge. (b) The targets with high UC receive more knowledge.

	SBP	DBP	BT	WBC	Results
1:00	100	53	40.1	12500	N/A
2:57	**89**	**46**	**N/A**	**N/A**	***Culture (+)***
5:00	120	64	N/A	N/A	N/A

	SBP	DBP	Temp	FiO2	Lactate	HCO_3^-	BUN	Cr
7:38	**N/A**	**37**	**37**	**0.35**	**5.3**	**N/A**	**N/A**	**N/A**
8:38	140	55	36.6	N/A	N/A	**10.6**	**85**	**4.2**
9:38	142	42	N/A	0.4	6.1	N/A	N/A	N/A

Table 4: Clinical Events in selected medical records for case studies. SBP - Systolic arterial blood pressure, DBP - Diastolic arterial blood pressure, BT - Body Temperature, WBC - White Blood Cell Count, FiO2 - Fractional inspired Oxygen, BUN - Blood Urine Nitrogen, Cr - Creatinine, Culture (+) - Blood Culture positive for *Klebsiella Pneumoniae*.



(a) MIMIC-III Infection

(b) PhysioNet

Figure 5: Visualizations of the amount of UC and normalized KT for example cases where the trends of both UC and KT at certain timesteps are correlated with noticeable clinical events (indicated with dotted arrow).

are not perfectly correlated, since the amount of knowledge transfer is also affected by the pairwise similarities between the source and the target features as well.

Interpretations of the Transfer Weights With the help of a physician, we further analyze how transfer weights and uncertainty are related with the patient's actual medical conditions (see Table 4 and Figure 5). We first consider an example record of a patient from the MIMIC-III Infection dataset who was suspected of infection on admission, and initially had fever, which was confirmed to be the symptom of bacterial infection later. Figure 5a shows the amount of knowledge transfer from task *Fever* at 3:00 to all later timesteps of task *Infection*. At this timestep, the patient's condition changes significantly. We observe that the patient had a fever, and the WBC level has increased to the state of leukocytosis, and both the SBP and DBP decrease over time. Most importantly, the patient is diagnosed to have an infection, as the culture results turns out to be positive for *Klebsiella Pneumoniae* at 2:57. With the drop of uncertainty of the task *Infection* around the time window where the event happens (dotted arrow in Figure 5a), the amount of knowledge transfer from *Fever* to *Infection* drops as well, as the knowledge from the source task becomes less useful.

As for another case study, we consider a record of a patient from PhyisoNet dataset who recovered from cardiac surgery and passed away during admission (Table 4 and Figure 5b). From Table 4, we observe that sign of multi-organ failure develops, as features related to respiratory (FiO_2, HCO_3^-, Lactate), renal (BUN, Creatinine), and cardiac (DBP) function deteriorates. As patient's condition after surgery gets worse, uncertainty of *Cardiac* starts to decrease at later timesteps (dotted arrow in Figure 5b) and knowledge transfer from *Cardiac* to *Mortality* increases as the uncertainty of the source task *Cardiac* starts to drop, since the knowledge from the source task becomes more reliable. Therefore, by analyzing the learned knowledge graph using our model, we could identify timesteps where meaningful interactions occur between tasks. For more case study examples of **MIMIC-III Heart Failure** and **Respiratory Failure** datasets, please see the **supplementary file**.

Model	Fever	Infection	Mortality	Average
AMTL-notransfer	0.7048 ± 0.01	0.6889 ± 0.01	0.6969 ± 0.01	0.6968 ± 0.01
AMTL-intratask	$\mathbf{0.7082 \pm 0.01}$	$\mathbf{0.7071 \pm 0.01}$	0.6814 ± 0.03	0.6989 ± 0.01
AMTL-samestep	$\mathbf{0.7130 \pm 0.00}$	0.6946 ± 0.01	0.6983 ± 0.03	$\mathbf{0.7019 \pm 0.01}$
TD-AMTL	0.6636 ± 0.03	0.6874 ± 0.01	0.6953 ± 0.02	0.6821 ± 0.00
TP-AMTL (constrained)	$\mathbf{0.7159 \pm 0.00}$	$\mathbf{0.7003 \pm 0.01}$	0.6561 ± 0.01	0.6908 ± 0.00
TP-AMTL (epistemic)	0.6997 ± 0.01	$\mathbf{0.7196 \pm 0.00}$	$\mathbf{0.7106 \pm 0.02}$	$\mathbf{0.7100 \pm 0.01}$
TP-AMTL (aleatoric)	0.6984 ± 0.02	$\mathbf{0.7032 \pm 0.01}$	$\mathbf{0.7217 \pm 0.02}$	0.7078 ± 0.00
TP-AMTL (full model)	$\mathbf{0.7165 \pm 0.00}$	$\mathbf{0.7093 \pm 0.01}$	$\mathbf{0.7098 \pm 0.01}$	$\mathbf{0.7119 \pm 0.00}$

Table 5: Ablation Study result on the MIMIC-III Infection Dataset.

Model	Stay 3	Cardiac	Recovery	Mortality	Average
AMTL-notransfer	$\mathbf{0.8901 \pm 0.01}$	0.9212 ± 0.01	$\mathbf{0.8986 \pm 0.01}$	$\mathbf{0.7515 \pm 0.01}$	$\mathbf{0.8653 \pm 0.01}$
AMTL-intratask	0.8829 ± 0.01	$\mathbf{0.9338 \pm 0.01}$	0.8812 ± 0.01	$\mathbf{0.7521 \pm 0.01}$	0.8625 ± 0.01
AMTL-samestep	0.8669 ± 0.01	0.9273 ± 0.01	0.8902 ± 0.01	0.7382 ± 0.01	0.8557 ± 0.01
TD-AMTL	0.7381 ± 0.06	0.9155 ± 0.01	0.8629 ± 0.01	0.7365 ± 0.01	0.8133 ± 0.02
TP-AMTL (constrained)	$\mathbf{0.8999 \pm 0.01}$	0.9186 ± 0.01	0.8892 ± 0.01	$\mathbf{0.7610 \pm 0.01}$	$\mathbf{0.8672 \pm 0.00}$
TP-AMTL (epistemic)	$\mathbf{0.8952 \pm 0.01}$	$\mathbf{0.9341 \pm 0.01}$	$\mathbf{0.8934 \pm 0.01}$	$\mathbf{0.7547 \pm 0.01}$	$\mathbf{0.8693 \pm 0.01}$
TP-AMTL (aleatoric)	0.8012 ± 0.03	0.9183 ± 0.01	0.8537 ± 0.02	0.7401 ± 0.03	0.8283 ± 0.01
TP-AMTL (full model)	$\mathbf{0.8953 \pm 0.01}$	$\mathbf{0.9416 \pm 0.01}$	$\mathbf{0.9016 \pm 0.01}$	$\mathbf{0.7586 \pm 0.01}$	$\mathbf{0.8743 \pm 0.01}$

Table 6: Ablation Study result on the PhysioNet Dataset.

Ablation Study

We compare our model against several variations of our model with varying knowledge transfer direction with respect to task (inter-task) and time (inter-timestep), and with temporal constraint (only future-to-past). Also we examine two kinds of uncertainty (epistemic, aleatoric) with our model (Table 5, 6).

Inter-Task and Inter-Timestep Knowledge Transfer.
1) AMTL-notransfer: The variant of our model without knowledge transfer.
2) AMTL-intratask: The variant of our model that knowledge only transfers within a same task.
3) AMTL-samestep: Another variant of our model that knowledge transfers only within a same time-step.
4) TD-AMTL: The deterministic counterpart of our model.

Our model outperforms **AMTL-intratask** and **AMTL-samestep**, which demonstrates the effectiveness of inter-task and inter-step knowledge transfer (Table 5, 6). **TD-AMTL** largely underperforms any variants, which may be due to overfitting of the knowledge transfer model, that can be effectively prevented by our bayesian framework.

Future-to-Past Transfer.
5) TP-AMTL (constrained): the model with temporal constraint

The unconstrained model outperforms **TP-AMTL (constrained)** (Table 5, 6), where transfer can only happen from the later timestep to earlier ones.

Two Kinds of Uncertainty.
6) TP-AMTL (epistemic) uses only MC-dropout to model epistemic uncertainty and $p_\theta(\mathbf{z}_d|\mathbf{x}, \boldsymbol{\omega})$ is simplified into $\mathcal{N}(\mathbf{z}_d; \boldsymbol{\mu}_d, \mathbf{0})$ (i.e. its pdf becomes the dirac delta function at $\boldsymbol{\mu}_d$ and $\mathbf{z}_d$ is always $\boldsymbol{\mu}_d$)

7) TP-AMTL (aleatoric) uses only $p_\theta(\mathbf{z}_d|\mathbf{x}, \boldsymbol{\omega})$ to model the aleatoric uncertainty, without MC-dropout.

For both MIMIC-III and PhysioNet datasets, epistemic uncertainty attributes more to the performance gain (Table 5, 6). However, it should be noted that the impacts of two kinds of uncertainty vary from dataset to dataset. By modelling both kinds of uncertainty, the model is guaranteed to get the best performance.

Conclusion

We propose a novel probabilistic asymmetric multi-task learning framework that allows asymmetric knowledge transfer between tasks at different timesteps, based on the uncertainty. While existing asymmetric multi-task learning methods consider asymmetric relationships between tasks as fixed, the task relationship may change at different timesteps in time-series data. Moreover, knowledge obtained for a task at a specific timestep could be useful for other tasks in later timesteps. To model the varying direction of knowledge transfer across tasks and timesteps, we propose a novel probabilistic multi-task learning framework that performs knowledge transfer based on the uncertainty of the latent representations for each task and timestep. We validate our model on four clinical time-series prediction tasks, on which our model shows strong performance over the baseline symmetric and asymmetric multi-task learning models without any sign of negative transfer. Case studies with learned knowledge graphs show that our model is interpretable, providing useful and reliable information on model predictions. This interpretability of our model will be useful in building a safe time-series analysis system for large-scale settings where both the number of time-series data instances and timestep are extremely large, such that manual analysis is impractical.

Acknowledgements

This work was supported by the Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2017-0-01779, A machine learning and statistical inference framework for explainable artificial intelligence(XAI)), the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921), and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3): 243–272.

Caruana, R. 1997. Multitask learning. *Machine learning* 28(1): 41–75.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8(1): 1–12.

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 *Advances in Neural Information Processing Systems*, 3504–3512.

Citi, L.; and Barbieri, R. 2012. PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. In *2012 Computing in Cardiology*, 257–260. IEEE.

Duong, L.; Cohn, T.; Bird, S.; and Cook, P. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, 845–850.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6(1): 1–18.

Heo, J.; Lee, H. B.; Kim, S.; Lee, J.; Kim, K. J.; Yang, E.; and Hwang, S. J. 2018. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*, 909–918.

Johnson, A. E.; Pollard, T. J.; and Mark, R. G. 2017. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, 361–376.

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* 3: 160035.

Kang, Z.; Grauman, K.; and Sha, F. 2011. Learning with Whom to Share in Multi-task Feature Learning. In *ICML*, volume 2, 4.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7482–7491.

Kumar, A.; and Daume III, H. 2012. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417* .

LeCun, Y.; and Cortes, C. 2010. *MNIST handwritten digit database*. hhttps://www.bibsonomy.org/bibtex/2935bad99fa1f65e03c25b315aa3c1032/mhwombat [Accessed: 03/01/2020].

Lee, G.; Yang, E.; and Hwang, S. 2016. Asymmetric multi-task learning based on task relatedness and loss. In *International Conference on Machine Learning*, 230–238.

Lee, H. B.; Yang, E.; and Hwang, S. J. 2017. Deep Asymmetric Multi-task Feature Learning. *arXiv preprint arXiv:1708.00260* .

Ma, L.; Gao, J.; Wang, Y.; Zhang, C.; Wang, J.; Ruan, W.; Tang, W.; Gao, X.; and Ma, X. 2020. AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 825–832.

Maurer, A.; Pontil, M.; and Romera-Paredes, B. 2013. Sparse coding for multitask and transfer learning. In *International conference on machine learning*, 343–351.

Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3994–4003.

Pirracchio, R. 2016. Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sicula) project. In *Secondary Analysis of Electronic Health Records*, 295–313. Springer.

Purushotham, S.; Meng, C.; Che, Z.; and Liu, Y. 2017. Benchmark of deep learning models on large healthcare mimic datasets. *arXiv preprint arXiv:1710.08531* .

Ruder, S.; Bingel, J.; Augenstein, I.; and Søgaard, A. 2017. Learning what to share between loosely related tasks. *ArXiv* .

Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

UdeM. 2014. Variations on the MNIST Digits. https://sites.google.com/a/lisa.iro.umontreal.ca/public_static_twiki/variations-on-the-mnist-digits [Accessed: 03/01/2020].

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Yang, Y.; and Hospedales, T. 2016a. Deep multi-task representation learning: A tensor factorisation approach. *arXiv preprint arXiv:1605.06391* .

Yang, Y.; and Hospedales, T. M. 2016b. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038* .