VARIATIONAL MULTI-TASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-task learning aims to improve the overall performance of a set of tasks by leveraging their relatedness. When training data is limited using priors is pivotal, but currently this is done in ad-hoc ways. In this paper, we develop *variational multi-task learning* - VMTL, a general probabilistic inference framework for simultaneously learning multiple related tasks. We cast multi-task learning as a variational Bayesian inference problem, which enables task relatedness to be explored in a principled way by specifying priors. We introduce Gumbel-softmax priors to condition the prior of each task on related tasks. Each prior is represented as a mixture of variational posteriors of other related tasks and the mixing weights are learned in a data-driven manner for each individual task. The posteriors over representations and classifiers are inferred jointly for all tasks and individual tasks are able to improve their performance by using the shared inductive bias. Experimental results demonstrate that VMTL is able to tackle challenging multi-task learning with limited training data well, and it achieves state-of-the-art performance on three benchmarks, consistently surpassing previous methods.

1 INTRODUCTION

Multi-task learning (Caruana, 1997) is a fundamental learning paradigm for machine learning, which aims to simultaneously solve multiple related tasks to improve the performance of the individual tasks by sharing knowledge. The crux of multi-task learning is how to explore task relatedness (Argyriou et al., 2007; Zhang & Yeung, 2012), which is non-trivial since the underlying relationship among tasks can be complicated and highly nonlinear. This has been extensively investigated in previous work by learning shared features, designing regularizers imposed on parameters (Pong et al., 2010; Kang et al., 2011; Jawanpuria et al., 2015; Jalali et al., 2010) or exploring priors over parameters (Heskes, 2000; Bakker & Heskes, 2003; Xue et al., 2007; Zhang & Yeung, 2012; Long et al., 2017). Recently, deep neural networks have been developed learning shared representations in the feature layers while keeping the classifier layers independent (Yang & Hospedales, 2016; Hashimoto et al., 2016; Ruder et al., 2017). It would be beneficial to learn them jointly by fully leveraging the shared knowledge related tasks, which however remains an open problem.

In our work, we consider a particularly challenging setting, where each task contains limited training data. Even more challenging, we have only a handful of related tasks to gain shared knowledge from. This is in stark contrast to few-shot learning (Gordon et al., 2018; Finn et al., 2017; Vinyals et al., 2016) that also suffers from limited data for each task, but usually having a large amount of related tasks. Therefore, in our scenario it is difficult to learn a proper model for each task independently without overfitting (Long et al., 2017; Zhang & Yang, 2017) and it is crucial to leverage the inductive bias (Baxter, 2000) provided by various other related tasks that are learned simultaneously. To do so, we employ the Bayesian framework as it is able to deliver uncertainty estimates on predictions and automatic model regularization (MacKay, 1992; Graves, 2011), which makes it well suited for multi-task learning with limited training data.

In this paper, we introduce *variational multi-task learning* - VMTL, a novel variational Bayesian inference approach that can explore task relatedness in a principled way. In order to fully utilize the shared knowledge from related tasks, we explore task relationships in both the feature representation and the classifier by placing prior distributions over them in a Bayesian framework. Thus, multi-task learning is cast as a variational inference problem for feature representations and classifiers jointly. To further leverage the shared knowledge from related tasks, we introduce the Gumbel-softmax prior for each task, which is a mixture of variational posteriors of other related tasks. We apply the

optimization technique (Jang et al., 2016) to learn the mixing weights jointly with the probabilistic modelling parameters by back-propagation. The Gumbel-softmax priors are incorporated into the inference of posteriors over representations and classifiers, which enable them to leverage the shared knowledge. We validate the effectiveness of the proposed VMTL by extensive evaluation on three challenging benchmarks for multi-task learning. The results demonstrate the benefit of variational Bayesian inference for modeling multi-task learning. VMTL consistently outperforms previous methods in terms of average accuracy of all tasks.

2 METHODOLOGY

In this work we tackle the challenging multi-task learning setting where only a few training samples are available for each task, and only a limited number of related tasks to share knowledge. We investigate multi-task learning under the Bayesian learning framework, where we learn the task relationship in a principled way by exploring priors. We cast multi-task learning as a variational inference problem, which offers a unified framework to learn task relatedness in both feature representations and task-specific classifiers.

2.1 MULTI-TASK VARIATIONAL INFERENCE

In our setting the tasks are classification problems which share the same label space, but where the samples are drawn from different data distributions. Given a set of related tasks $\{\mathcal{D}_t\}_{t=1}^{T}$ and each task $\mathcal{D}_t = \{\mathbf{x}_t^n, \mathbf{y}_t^n\}_{n=1}^{N_t}$, N_t is the number of training samples in the t-th task, the goal under this setting is to predict the label $\mathbf{y}$ of the test sample $\mathbf{x}$ for all tasks simultaneously, using the shared information extracted from other related tasks. We note that the main challenge is the limited number of labeled samples for each task, which makes it difficult to learn a proper model for each tasks independently (Long et al., 2017; Zhang & Yang, 2017).

Under this multi-task learning setting, we consider the Bayesian treatment. For a single task without knowledge sharing from related takes, we place a prior over its classifier parameter $\mathbf{w}$, which gives rise to the following data log-likelihood to maximize:

$$\log p(\mathcal{D}) = \log \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{1}$$

For multi-task learning, we solve T tasks simultaneously with knowledge sharing among tasks. Thus, after observing data from all T tasks, the true posterior $p(\mathbf{w}_t|\mathcal{D}_t)$ of a single task t becomes $p(\mathbf{w}_t|\mathcal{D}_{1:T})$. Using Bayes' rule, we have the posterior for task t as follows:

$$p(\mathbf{w}_t|\mathcal{D}_{1:T}) \propto p(\mathbf{w}_t) \prod_{t=1}^{T} p(\mathcal{D}_t|\mathbf{w}_t) \propto p(\mathbf{w}_t|\mathcal{D}_{1:T} \backslash \mathcal{D}_t) p(\mathcal{D}_t|\mathbf{w}_t). \tag{2}$$

We introduce a variational distribution $q(\mathbf{w}_t; \theta_t)$ parameterized by θ_t for current task t to approximate the true posterior, which involves minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior:

$${\theta_t}^* = \arg\min_{\theta_t} \mathbb{D}_{\mathrm{KL}} \Big[q(\mathbf{w}_t; \theta_t) || p(\mathbf{w}_t|\mathcal{D}_{1:T}) \Big]. \tag{3}$$

Generally, the approximate posterior is defined as a fully-factorized Gaussian distribution, i.e. $q(\mathbf{w}_t; \theta_t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2)$ (Nguyen et al., 2017; Kingma et al., 2015; Molchanov et al., 2017). By applying Eq. (2) into (3) and extending them to all tasks, we obtain an evidence lower bound (ELBO) for multi-task learning:

$$\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t}^{T} \Big[\mathbb{E}_{\mathbf{w}_t \sim q} \Big[\log p(\mathcal{D}_t|\mathbf{w}_t) \Big] - \mathbb{D}_{\mathrm{KL}} \Big[q(\mathbf{w}_t; \theta_t) || p(\mathbf{w}_t|\mathcal{D}_{1:T} \backslash \mathcal{D}_t) \Big] \Big], \tag{4}$$

where $\boldsymbol{\theta} = \{\theta_t\}_{t=1}^{T}$ is the set of parameters for all task-specific classifiers. We maximize the ELBO to optimize the model parameters for multi-task learning. It is worth noting that for MTL the prior of

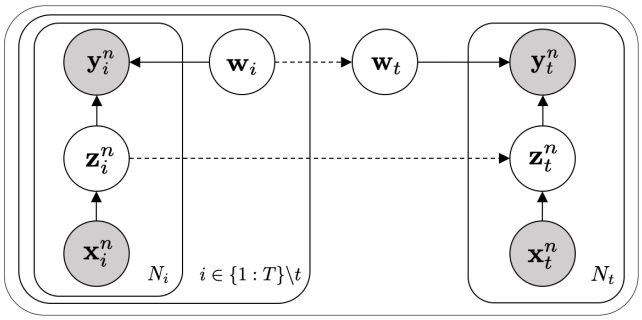

Fig. 1. A graphical illustration of variational multi-task learning (VMTL). VMTL offers a principled way to explore task relationships: for each task, the priors over the classifiers and feature representations are conditioned on other tasks.

each task is conditioned on other tasks which allows knowledge sharing between them, in contrast to single-task learning without knowledge sharing where uninformative Gaussian priors are generally applied. Actually, this multi-task ELBO in Eq. (4) provides a general probabilistic inference framework that enables task relatedness to be explored in a principled way by leveraging the inductive bias provided by other related tasks (Ruder, 2017). For each task, the conditional prior $p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)$ will serve as a regularizer for the posterior inference.

In order to fully leverage the shared knowledge from related tasks to improve each individual task, in addition to the classifiers, we would also like to share knowledge among the feature representations of samples from different tasks. To this end, we introduce the conditional prior $p(\mathbf{z}_t^n|\mathbf{x}_t^n)$ over the feature representation $\mathbf{z}_t^n$ for each sample $\mathbf{x}_t^n$ in task t. In doing so, we are able to explore task relatedness among feature representations in a unified way, as we do for classifiers.

To this end, we rewrite the log-likelihood as a sum over the marginal likelihoods of individual data points as follows:

$$\log p(\mathcal{D}_t|\mathbf{w}_t) = \frac{1}{N_t}\sum_{n}^{N_t} \log p(\mathbf{y}_t^n|\mathbf{w}_t, \mathbf{x}_t^n). \tag{5}$$

The variational Bayesian inference framework developed in Eq. (4) allows the latent representation to be seamlessly incorporated into the log-likelihood in Eq. (5). Under the assumption that $\mathbf{w}_t$ and $\mathbf{z}_t^n$ are conditionally independent, we therefore obtain a new marginal conditional log-likelihood as follows:

$$\log p(\mathbf{y}_t^n|\mathbf{w}_t, \mathbf{x}_t^n) = \log \int p(\mathbf{y}_t^n, \mathbf{z}_t^n|\mathbf{w}_t, \mathbf{x}_t^n) d\mathbf{z}_t^n = \log \int p(\mathbf{y}_t^n|\mathbf{z}_t^n, \mathbf{w}_t) p(\mathbf{z}_t^n|\mathbf{x}_t^n) d\mathbf{z}_t^n. \tag{6}$$

The posterior $p(\mathbf{z}_t^n|\mathbf{x}_t^n, \mathbf{y}_t^n)$ of latent representations $\mathbf{z}_t$ is intractable. Therefore, like VAEs (Kingma & Welling, 2013) we introduce again a variational posterior namely $q(\mathbf{z}_t|\mathbf{x}_t;\phi)$, which is made conditional on the sample $\mathbf{x}_t^n$. We note that ϕ is the parameters of the inference network for latent representations, which is optimized jointly with other parameters in our model. By introducing $q(\mathbf{z}_t^n|\mathbf{x}_t^n;\phi)$ into Eq. (6) and applying Jensen's inequality, we have

$$\log p(\mathbf{y}_t^n|\mathbf{w}_t, \mathbf{x}_t^n) \geq \mathbb{E}_{\mathbf{z}_t^n \sim q_{\mathbf{z}_t^n}}\Big[\log p(\mathbf{y}_t^n|\mathbf{z}_t^n, \mathbf{w}_t)\Big] - \mathbb{D}_{\mathrm{KL}}\Big[q(\mathbf{z}_t^n|\mathbf{x}_t^n;\phi)||p(\mathbf{z}_t^n|\mathbf{x}_t^n)\Big]. \tag{7}$$

After integrating Eq. (4) and (7), we obtain the following variational objective for multi-task learning:

$$\begin{aligned}\mathcal{L}_{\mathrm{VMTL}}(\boldsymbol{\theta}, \phi) = \frac{1}{T}\sum_{t}^{T}\Bigg[\frac{1}{N_t}\sum_{n}^{N_t}\Big[\mathbb{E}_{q_{\mathbf{w}_t}}\mathbb{E}_{q_{\mathbf{z}_t^n}}\big[\log p(\mathbf{y}_t^n|\mathbf{z}_t^n, \mathbf{w}_t)\big] - \mathbb{D}_{\mathrm{KL}}\big[q(\mathbf{z}_t^n|\mathbf{x}_t^n;\phi)||p(\mathbf{z}_t^n|\mathbf{x}_t^n)\big]\Big] \\ - \mathbb{D}_{\mathrm{KL}}\big[q(\mathbf{w}_t;\theta_t)||p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)\big]\Bigg].\end{aligned} \tag{8}$$

The objective function provides a general probabilistic inference framework for multi-task learning, which allows us to jointly explore shared knowledge among representations and classifiers in a

unified way by specifying priors. The detailed derivation of the evidence lower bound with $\mathbf{z}$ and $\mathbf{w}$ jointly is given in the Appendix A.1. The graphical illustration of our variational multi-task learning is shown in Fig. 1.

2.2 Learning Task Relatedness via Gumbel-softmax Priors

The proposed variational multi-task inference framework enables the relationship among tasks to be explored by designing priors for both latent representations and classifiers. we introduce the Gumbel-softmax prior for each task, which is a mixture of variational posteriors of other related works. The learnable mixing weight indicates the relevance of a related task to the current task. In the case of latent representations, we define the prior by using other tasks posteriors of latent representations:

$$p(\mathbf{z}_t|\mathbf{x}_t) := \sum_{i\in\{1,\cdots,T\}\setminus t} \mathcal{A}_{ti} q(\mathbf{z}_i|\mathbf{m}_i;\phi), \tag{9}$$

where $\mathbf{m}_i$ is the mean feature representation of samples from the same class in the i-th task. We note that $\mathcal{A}_{ti}$ is the mixing weight. We apply the Gumbel-softmax technique (Jang et al., 2016) to learn the mixing weight between any pair of tasks with the probabilistic modelling parameters through back-propagation:

$$\mathcal{A}_{ti} = \frac{\exp((\log \pi_{ti} + g_{ti})/\tau)}{\exp((\log \pi_{ti} + g_{ti})/\tau) + \exp((\log(1-\pi_{ti}) + g'_{ti})/\tau)}, \tag{10}$$

where g_{ti} and g'_{ti} are samples taken from a Gumbel distribution, using inverse transform sampling by drawing $u \sim \text{Uniform}(0,1)$ and computing $g = -\log(-\log(u))$. Parameter τ is the softmax temperature.

Likewise, we specify the prior over classifier parameters $\mathbf{w}_t$ in the same way as in $\mathbf{z}_t$:

$$p(\mathbf{w}_t|\mathcal{D}_{1:T}\setminus\mathcal{D}_t) := \sum_{i\in\{1,\cdots,T\}\setminus t} \mathcal{A}^{(\mathbf{w})}_{ti} q(\mathbf{w}_i;\theta_i), \tag{11}$$

where $\mathcal{A}^{(\mathbf{w})}_{ti}$ is obtained in a similar way as in Eq. (10). Note that we use different mixing weights in designing the priors for representations and classifiers, which gave better results than a shared one in our preliminary experiments. This is likely due to the fact that the representations and classifiers leverage different correlation patterns among different tasks.

2.3 Amortized Inference

We leverage an amortization technique (Gershman & Goodman, 2014), in which we amortize the computational cost of inferring the posterior of the latent representation, as done in VAEs (Kingma & Welling, 2013). In effect, the amortized inference can also be adopted to learn classifier parameters similar to the probabilistic prediction in few-shot learning (Gordon et al., 2018).

To this end, we design the variational posterior $q(\mathbf{w}_t|\mathcal{D}_t)$ in a context-independent manner such that each weight vector $\mathbf{w}^c_t$ depends only on samples from class c of the current task t:

$$q(\mathbf{w}_t|\mathcal{D}_t) = \prod_{c=1}^{C} q(\mathbf{w}^c_t|\mathcal{D}^c_t) = \prod_{c=1}^{C} \mathcal{N}(\boldsymbol{\mu}^c_t, \text{diag}((\boldsymbol{\sigma}^c_t)^2)), \tag{12}$$

where $\mathcal{D}^c$ are the samples from the c-th class, C is the size of the shared label space, and each posterior is parameterized as a diagonal Gaussian distribution. The amortized inference of these posteriors is implemented by multi-layer perceptrons (MLPs) (Gordon et al., 2018; Edwards & Storkey, 2016; Kingma & Welling, 2013) and the parameters of the inference networks are jointly optimized in the end-to-end learning. In practice, we take the mean of feature representations in each class as input and return the parameters μ^c_t and σ^c_t for this class. In contrast to the amortized inference for latent representations, the amortized classifier inference shares the cost across classes and reduces the overall cost. It offers an effective way to handle a huge number of classes and produces competitive performance even in existence of the amortization gap (Cremer et al., 2018).

2.4 EMPIRICAL OBJECTIVE FUNCTION

In our implementation for the variational objective for multi-task learning Eq. (8), we adopt Monte Carlo sampling, and obtain the empirical objective function as follows:

$$\begin{aligned}\hat{\mathcal{L}}_{\text{VMTL}}(\boldsymbol{\theta}, \phi) = \frac{1}{T}\sum_{t}^{T}\Bigg[\frac{1}{N_t}\sum_{n}^{N_t}\Big[\frac{1}{ML}\sum_{\ell=1}^{L}\sum_{m=1}^{M}\big[\log p(\mathbf{y}_t|\mathbf{z}_t^{n,(\ell)}, \mathbf{w}_t^{(m)})\big] \\ - \mathbb{D}_{\text{KL}}\big[q(\mathbf{z}_t^n|\mathbf{x}_t^n;\phi)||p(\mathbf{z}_t^n|\mathbf{x}_t^n)\big]\Bigg] - \mathbb{D}_{\text{KL}}\big[q(\mathbf{w}_t;\theta_t)||p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)\big]\Bigg],\end{aligned} \tag{13}$$

where $\mathbf{z}_t^{n,(\ell)} \sim q(\mathbf{z}_t^n|\mathbf{x}_t^n;\phi)$, $\mathbf{w}_t^{(m)} \sim q(\mathbf{w}_t;\theta)$. L and M are the number of Monte Carlo samples. We maximize this empirical function to optimize the model's parameters: the log-likelihood term is implemented as the cross-entropy loss and the KL terms can be computed in closed forms. To sample from the variational posteriors, we adopt the reparameterization trick (Kingma & Welling, 2013). In the posterior inference of classifiers without amortization, we use the local reparameterization trick to reduce the gradient variance (Kingma et al., 2015). The priors $p(\mathbf{z}_t^n|\mathbf{x}_t^n)$ and $p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)$ are implemented with Gumbel-softmax priors provided in Eq. (9) and Eq. (11), respectively. For amortized classifiers, the variational posterior $q(\mathbf{w}_t;\theta)$ is implemented using Eq. (12).

3 RELATED WORKS

Multi-task learning (Caruana, 1997) is a machine learning paradigm that aims to leverage shared knowledge from multiple related tasks to improve the generalization performance of all the tasks simultaneously. The core of multi-task learning is how to explore the task relationship, which has been extensively investigated in the literature.

Early works design feature-based or parameter-based regularizations to explore task relationships (Liu et al., 2009; Obozinski et al., 2010; Pong et al., 2010; Jalali et al., 2010; Kang et al., 2011; Jawanpuria et al., 2015; Long et al., 2017; Zhang et al., 2020). Obozinski et al. (2010) are the first to study the multi-task feature selection (MTFS) problem based on the $l_{2,1}$ norm. Liu et al. (2009) propose to use the $l_{\infty,1}$ norm with the objective function to select important features. Kang et al. (2011) design multiple task clusters, aiming to minimize the squared trace norm of the classifier parameters in each cluster.

Bayesian methods (Heskes, 2000; Bakker & Heskes, 2003; Yu et al., 2005; Xue et al., 2007; Guo et al., 2011; Titsias & Lázaro-Gredilla, 2011; Zhang & Yeung, 2012; Long et al., 2017) are developed for multi-task learning under probabilistic frameworks, where the regularization usually corresponds to a prior. Heskes (2000) proposes a Bayesian neural network for multi-task learning, and analyse it with an empirical Bayesian framework. Yu et al. (2005) investigate Gaussian processes for multi-task learning assuming that all models are sampled from a common prior. Zhang & Yeung (2012); Long et al. (2017) reformulate the $l_{p,q}$ norm regularizer as a matrix-variate generalized normal prior and utilize the prior information to explore task relations. Further, Qian et al. (2020) adopt a variational information bottleneck method (Alemi et al., 2016) with an uninformative prior distribution to improve the latent probabilistic representation. An important conclusion drawn by (Qian et al., 2020) is that, under adversarial attacks, variational latent representations are regularized and thereby expected to be more robust to noise than deterministic latent representations. Although these methods achieve encouraging improvements, the performance remains unsatisfactory when the number of training data is very limited.

Deep learning has recently been explored for multi-task learning (Misra et al., 2016; Yang & Hospedales, 2016; Hashimoto et al., 2016; Ruder et al., 2017; Long et al., 2017; Meyerson & Miikkulainen, 2017; Chen et al., 2018; Maziarz et al., 2019) by designing different deep architectures to explore task relationships. Liu et al. (2019); Zhang et al. (2020) use a hard parameter-sharing encoder to extract the shared representations, while learning a task-specific decoder to obtain the pixel-level predictions. Based on soft parameter sharing, Misra et al. (2016) propose cross-stitch units to allow the model to leverage the shared knowledge from another task. Gao et al. (2020) follow the soft parameter sharing mechanism and incorporate neural architecture search into general-purpose multi-task learning. Meyerson & Miikkulainen (2017); Rosenbaum et al. (2017); Maziarz et al. (2019); Strezoski et al. (2019) develop flexible soft-ordering approaches to enable more effec-

tive sharing among tasks. Instead of learning the structure of sharing, Kendall et al. (2018) propose to weight multiple loss functions by considering the homoscedastic uncertainty of each task. Chen et al. (2018) present a gradient normalization algorithm that automatically balances training in deep multi-task models by dynamically tuning gradient magnitudes.

In our work, we address multi-task learning in a probabilistic inference framework by casting it as a variational Bayesian inference problem. We explore task relationships in a principled way by specifying priors conditioned on other tasks, which enables the model to share knowledge among related tasks for learning both representations and classifiers.

4 EXPERIMENTS

We conduct experiments on three commonly-used benchmarks for multi-task learning with limited data. The results demonstrate the benefits of variational Bayesian approximation to representations and classifiers for multi-task learning. Our VMTL achieves the best performance and consistently surpasses previous methods. We provide more experimental results in the Appendix B.

4.1 DATASETS

We evaluate the proposed VMTL under a challenging multi-task learning setting, where each task has limited training data and only a handful of related tasks that can be learned simultaneously to leverage the shared knowledge. To benchmark our model with previous methods, we conduct experiments on three benchmark datasets, where tasks are defined as classification problems from distinctive domains with a shared label space.

Office-Home (Venkateswara et al., 2017) contains images from four domains/tasks: Artistic (A), Clipart (C), Product (P) and Real-world (R). Each task contains images from 65 object categories collected in office and home settings. There are about 15,500 images in total.

Office-Caltech (Gong et al., 2012) was constructed by selecting the ten categories shared between Office-31 (Saenko et al., 2010) and Caltech-256 datasets (Griffin et al., 2007). One task consists of data from Caltech-256 (C), and the other three tasks consist of data from Office-31 whose images were collected from three distinct domains/tasks, e.g., Amazon (A), Webcam (W) and DSLR (D). There are 8-151 samples per category per task, and 2,533 images in total.

ImageCLEF (Long et al., 2017), the benchmark for the ImageCLEF domain adaptation challenge, contains 12 common categories shared by four public datasets/tasks: Caltech-256 (C), ImageNet ILSVRC 2012 (I), Pascal VOC 2012 (P), and Bing (B). There are 2,400 images in total.

We follow the standard evaluation protocol (Zhang & Yeung, 2012; Zhang & Yang, 2017) for multi-task learning and randomly select 5%, 10%, and 20% of samples from each task as the training set and use the remaining samples as the test set (Long et al., 2017). Note that when we use 5%, 10%, and 20% labeled data for training, there are on average respectively 3, 6 and 12 samples per category per task. Hence, each task is provided with a limited amount of labeled data, which is insufficient for building reliable classifiers without overfitting.

4.2 EXPERIMENTAL SETTINGS

Implementation Details Following the experimental settings in (Long et al., 2017), we remove the final classifier layer of VGGnet (Simonyan & Zisserman, 2014) and apply the remaining model as the feature extractor in our architecture. In our experiments, the temperature is annealed using the same schedule as applied in (Jang et al., 2016). For the KL-divergence we use the annealing scheme from (Bowman et al., 2015), increasing the weight of the KL-divergence by a rate of 1e-6 per iteration. The dimension of the latent variable is set to 512. We adopt the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 1e-4 for training.

Comparison Methods We compare our method with single-task learning (STL) as well as a variational version of STL (VSTL) implemented by introducing variational Bayesian inference to representations and classifiers without using task relationship (Details can be found in the Appendix A.2). We also define a basic multi-task learning (bMTL) model, which is a deep learning model with a shared feature extractor and task-specific classifiers. The bMTL serves as a baseline model to

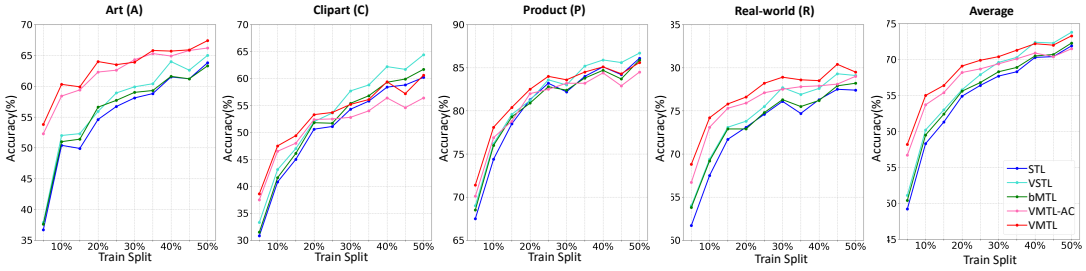

Fig. 2. The performance under different proportions of training data on the *Office-Home* dataset.

demonstrate the benefits of the probabilistic modeling of multi-task learning based on variational Bayesian inference. VMTL is our basic proposed method. VMTL-AC is our proposed method with amortized classifiers. We compare with multilinear relationship network (MRN) (Long et al., 2017), which holds the best performance among previous methods on the three benchmarks.

4.3 Experimental Results

Effectiveness in Handling Limited Data We conduct ablation studies to demonstrate the benefits of the proposed variational multi-task learning in exploring task relatedness. We conduct experiments under a large range of train-test splits from 5% to 50%. The results on the Office-Home dataset for different tasks and average accuracy of all tasks are illustrated in Fig. 2. More detailed experimental results are provided in the Appendix B.1. The performance advantage of our VMTL is larger in the settings with less training data ($\leq 20\%$), which demonstrates its effectiveness in handling challenging scenarios with very limited training data.

Effectiveness of Variational Bayesian Approximation We investigate the effect of variational Bayesian inference for representations and classifiers, separately. We conduct these experiments on the Office-Home dataset. The results with different train-test splits are shown in Table 1. More detailed experimental results are provided in the Appendix B.2. As can be seen, both variational Bayesian representations and classifiers can benefit the performance. This benefit becomes more significant when training data is very limited, which indicates the effectiveness of leveraging shared knowledge by conditioning priors on related tasks in our VMTL. In addition, in Fig. 2, the VSTL is shown to be a strong learning model, which again demonstrates the benefits of variational Bayesian representations and classifiers compared to STL.

Table 1. Effectiveness of variational Bayesian approximation for representations and classifiers.

Representation ($\mathbf{z}$)	Classifier ($\mathbf{w}$)	5%	10%	20%
×	×	50.4	59.5	65.6
✓	×	51.3	60.4	66.7
×	✓	56.8	63.7	68.5
✓	✓	**58.2**	**65.0**	**69.1**

Effectiveness of Gumbel-Softmax Priors The introduced Gumbel-softmax priors provide an effective way to learn data-driven task relationships. To demonstrate their effectiveness, we compare with several alternatives, including the mean and the learnable weighted posteriors of other tasks. The comparison results are shown in Table 2. The proposed Gumbel-softmax priors perform the best, consistently surpassing other alternatives. It is also worth noting that the advantage of Gumbel-softmax priors is ever larger for very limited training data, e.g., 5%, a challenging scenario where it is crucial to leverage task relatedness.

Table 2. Performance comparison of VMTL with different priors. The detailed results on each individual task are provided in the Appendix B.3. Our Gumbel-softmax produces the best performance.

Priors	5%	10%	20%
Mean	57.2	64.5	68.6
Learnable weighted	57.0	64.4	68.6
Gumbel-softmax	**58.2**	**65.0**	**69.1**

Table 3. Performance comparison of different methods on the *Office-Home* dataset for multiple tasks: Artistic (A), Clipart (C), Product (P) and Real-world (R).

Methods	5%					10%					20%				
	A	C	P	R	Avg.	A	C	P	R	Avg.	A	C	P	R	Avg.
STL	36.7	30.8	67.5	61.7	49.2	50.4	40.8	74.4	67.5	58.3	54.6	50.6	81.3	73.1	64.9
VSTL	37.9	33.3	69.0	64.0	51.1	52.0	43.1	76.2	69.4	60.2	55.9	52.0	81.2	73.8	65.8
bMTL	37.6	31.5	68.5	63.8	50.4	51.0	41.6	76.0	69.2	59.5	56.6	51.8	80.9	72.9	65.6
MRN	53.3	36.4	70.5	67.7	57.0	59.9	42.7	76.3	73.0	63.0	58.5	55.6	80.7	72.8	66.9
VMTL-AC	52.3	37.5	70.1	66.7	56.7	58.4	46.5	76.9	73.1	63.7	62.3	52.4	82.0	75.9	68.2
VMTL	53.8	38.6	71.4	68.8	**58.2**	60.3	47.5	78.1	74.2	**65.0**	64.0	53.3	82.5	76.6	**69.1**

Table 4. Performance comparison of different methods on the *Office-Caltech* dataset for multiple tasks: Amazon (A), Webcam (W), DSLR (D) and Caltech-256 (C).

Methods	5%					10%					20%				
	A	W	D	C	Avg.	A	W	D	C	Avg.	A	W	D	C	Avg.
STL	87.4	87.9	96.4	82.8	88.6	92.8	97.7	87.8	84.3	90.7	94.9	92.8	95.2	86.7	92.4
VSTL	88.3	89.1	97.0	81.4	89.0	93.1	96.6	90.0	84.5	91.1	95.5	94.5	96.0	87.7	93.4
bMTL	90.0	89.4	95.0	83.5	89.5	93.6	97.0	92.1	86.3	92.3	95.0	94.5	96.0	86.8	93.1
MRN	92.7	94.3	97.1	89.2	93.4	95.0	98.1	95.0	91.3	94.8	95.5	94.9	99.2	91.0	95.1
VMTL-AC	93.2	95.0	96.1	89.7	93.5	94.8	96.8	97.7	90.1	94.9	95.2	95.7	99.4	90.7	95.2
VMTL	93.8	95.5	96.4	90.0	**93.9**	95.5	97.0	97.9	91.0	**95.3**	95.6	95.8	99.2	90.6	**95.3**

Comparison with other methods The comparison results on the Office-Home, Office-Caltech and ImageCLEF datasets are shown in Tables 3, 4 and 5, respectively. A comprehensive comparison with more other methods is provided in the Appendix B.4. Multi-task learning aims to improve the overall performance of all tasks simultaneously, thus the average accuracy of all tasks provides a proper evaluation for multi-task learning models. The best results of average accuracy are marked in bold, while the second best by underline.

Our VMTL consistently achieves the best performance on all datasets with all train-test split settings. It is worth highlighting that on the most challenging setting of 5% training data, our VMTL shows a large performance advantage over compared methods. Specifically, on the ImageCLEF dataset, under the 5% setting, our VMTL surpasses the second best by a phenomenal margin up to 2.3%. This demonstrates the effectiveness of VMTL in exploring relatedness to improve the performance of each task. In addition, our VMTL with amortized classifiers (VMTL-AC) can also produce comparable performance, and is better than most previous methods. Finally, the performance improvement of VSTL over STL also indicates the benefits of variational Bayesian approximation for representations and classifiers.

Table 5. Performance comparison of different methods on the *ImageCLEF* dataset for multiple tasks: Caltech-256 (C), ImageNet ILSVRC 2012 (I), Pascal VOC 2012 (P), and Bing (B).

Methods	5%					10%					20%				
	C	I	P	B	Avg.	C	I	P	B	Avg.	C	I	P	B	Avg.
STL	85.4	71.4	57.7	36.0	62.6	88.9	77.8	64.3	47.6	69.7	92.9	84.6	72.5	54.6	76.2
VSTL	87.0	73.2	60.5	39.0	64.9	89.6	79.1	66.6	48.0	70.8	93.3	87.3	72.7	55.4	77.2
bMTL	88.3	73.2	61.2	40.0	65.7	90.6	79.3	66.3	51.9	72.0	92.9	86.5	71.9	56.0	76.8
MRN	90.1	76.5	72.8	54.9	73.7	93.3	83.2	70.4	56.3	75.8	94.4	89.2	75.8	59.4	79.7
VMTL-AC	89.1	81.4	71.2	56.5	74.5	92.3	83.9	71.8	58.9	76.7	92.9	87.7	77.0	60.1	79.4
VMTL	91.1	83.2	71.4	58.3	**76.0**	93.7	86.5	71.8	59.5	**77.9**	94.0	89.7	77.9	59.7	**80.3**

5 Conclusion

In this paper, we address the multi-task learning problem and tackle a challenging setting where each task has a very limited amount of training data, with only a handful of related tasks. To this end, we develop *variational multi-task learning* - VMTL, a general probabilistic inference framework for simultaneously learning multiple tasks. We cast multi-task learning as a variational inference problem, which enables task relationships to be explored in a principled way by specifying priors. Specifically, we introduce the Gumbel-softmax priors, which offer an effective way to learn the task relatedness in a data-driven manner for each task. We evaluate VMTL on three widely used benchmark datasets for multi-task learning. Results demonstrate that our VMTL consistently achieves better or comparable performance with state-of-the-art multi-task learning approaches.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 5

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in neural information processing systems*, pp. 41–48, 2007. 1, 14

Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4(May):83–99, 2003. 1, 5

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000. 1

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015. 6

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 1, 5

Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 42–50, 2011. 14

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pp. 794–803. PMLR, 2018. 5, 6

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. *arXiv preprint arXiv:1801.03558*, 2018. 4

Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016. 4

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017. 1

Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11543–11552, 2020. 5

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014. 4

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073. IEEE, 2012. 6

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018. 1, 4

Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011. 1

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6

Shengbo Guo, Onno Zoeter, and Cédric Archambeau. Sparse bayesian multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 1755–1763, 2011. 5

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*, 2016. 1, 5

TM Heskes. Empirical bayes for learning to learn. 2000. 1, 5

Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in neural information processing systems*, pp. 964–972, 2010. 1, 5

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2, 4, 6

Pratik Kumar Jawanpuria, Maksim Lapin, Matthias Hein, and Bernt Schiele. Efficient output kernel learning for multiple tasks. In *Advances in neural information processing systems*, pp. 1189–1197, 2015. 1, 5

Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *ICML*, volume 2, pp. 4, 2011. 1, 5

Alex Kendall, Yarin Gal, and Roberto Cipolla. 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*, pp. 7482–7491, 2018. 6

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 4, 5

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pp. 2575–2583, 2015. 2, 5, 12

Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 649–656, 2009. 5

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1871–1880, 2019. 5

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and S Yu Philip. Learning multiple tasks with multilinear relationship networks. In *Advances in neural information processing systems*, pp. 1594–1603, 2017. 1, 2, 5, 6, 7, 14

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 1

Krzysztof Maziarz, Efi Kokiopoulou, Andrea Gesmundo, Luciano Sbaiz, Gabor Bartok, and Jesse Berent. Gumbel-matrix routing for flexible multi-task learning. *arXiv preprint arXiv:1910.04915*, 2019. 5

Elliot Meyerson and Risto Miikkulainen. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. *arXiv preprint arXiv:1711.00108*, 2017. 5

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4003, 2016. 5

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. *arXiv preprint arXiv:1701.05369*, 2017. 2, 12

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017. 2

Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010. 5

Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010. 1, 5

Weizhu Qian, Bowei Chen, and Franck Gechter. Multi-task variational information bottleneck. *arXiv preprint arXiv:2007.00339*, 2020. 5

Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017. 5

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 3

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Sluice networks: Learning what to share between loosely related tasks. *stat*, 1050:23, 2017. 1, 5

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010. 6

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pp. 3483–3491, 2015. 12

Gjorgji Strezoski, Nanne van Noord, and Marcel Worring. Many task learning with task routing. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1375–1384, 2019. 5

Michalis K Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in neural information processing systems*, pp. 2339–2347, 2011. 5

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016. 1

Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(Jan):35–63, 2007. 1, 5

Yongxin Yang and Timothy Hospedales. Deep multi-task representation learning: A tensor factorisation approach. *arXiv preprint arXiv:1605.06391*, 2016. 1, 5, 14

Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd international conference on Machine learning*, pp. 1012–1019, 2005. 5

Yi Zhang, Yu Zhang, and Wei Wang. Deep multi-task learning via generalized tensor trace norm. *arXiv preprint arXiv:2002.04799*, 2020. 5

Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017. 1, 2, 6

Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012. 1, 5, 6, 14

A DERIVATION

A.1 DERIVATION OF EVIDENCE LOWER BOUND FOR MULTI-TASK LEARNING

We provide derivation of the evidence lower bound with $\mathbf{z}$ and $\mathbf{w}$ jointly. The log likelihood of task t is conditioned on the data from other related tasks.

$$\log p(\mathbf{y}_t|\mathbf{x}_t, \mathcal{D}_{1:T}\backslash\mathcal{D}_t) = \log \int\int p(\mathbf{y}_t, \mathbf{z}_t, \mathbf{w}_t, |\mathbf{x}_t, \mathcal{D}_{1:T}\backslash\mathcal{D}_t) d\mathbf{w}_t d\mathbf{z}_t \tag{14}$$

Under the assumption that $\mathbf{w}_t$ and $\mathbf{z}_t$ are conditionally independent, we therefore obtain

$$\begin{aligned}
&\log p(\mathbf{y}_t|\mathbf{x}_t, \mathcal{D}_{1:T}\backslash\mathcal{D}_t) \\
&= \log \int\int p(\mathbf{y}_t|\mathbf{z}_t, \mathbf{w}_t) p(\mathbf{z}_t|\mathbf{x}_t) p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t) d\mathbf{w}_t d\mathbf{z}_t \\
&= \log \int \big[\int p(\mathbf{y}_t|\mathbf{z}_t, \mathbf{w}_t) p(\mathbf{z}_t|\mathbf{x}_t) d\mathbf{z}_t \big] p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t) d\mathbf{w}_t \\
&= \log \int \frac{\big[\int p(\mathbf{y}_t|\mathbf{z}_t, \mathbf{w}_t) p(\mathbf{z}_t|\mathbf{x}_t) d\mathbf{z}_t\big] p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t) q(\mathbf{w}_t)}{q(\mathbf{w}_t)} d\mathbf{w}_t \\
&\geq -\mathbb{KL}[q(\mathbf{w}_t)||p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)] + \mathbb{E}_{q(\mathbf{w}_t)}\big[\log \int p(\mathbf{y}_t|\mathbf{z}_t, \mathbf{w}_t) p(\mathbf{z}_t|\mathbf{x}_t) d\mathbf{z}_t\big] \\
&\geq -\mathbb{KL}[q(\mathbf{w}_t)||p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)] + \mathbb{E}_{q(\mathbf{w}_t)}\big[\log \int \frac{p(\mathbf{y}_t|\mathbf{z}_t, \mathbf{w}_t) p(\mathbf{z}_t|\mathbf{x}_t) q(\mathbf{z}_t|\mathbf{x}_t)}{q(\mathbf{z}_t|\mathbf{x}_t)} d\mathbf{z}_t\big] \\
&\geq -\mathbb{KL}[q(\mathbf{w}_t)||p(\mathbf{w}_t|\mathcal{D}_{1:T}\backslash\mathcal{D}_t)] - \mathbb{KL}[q(\mathbf{z}_t|\mathbf{x}_t)||p(\mathbf{z}_t|\mathbf{x}_t)] + \mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{y}_t|\mathbf{z}_t, \mathbf{w}_t)]
\end{aligned} \tag{15}$$

A.2 DERIVATION OF EVIDENCE LOWER BOUND FOR SINGLE-TASK LEARNING

Generally, the proposed Bayesian inference framework which infers the posteriors of presentations $\mathbf{z}$ and classifiers $\mathbf{w}$ jointly can be widely applied in other research fields. To be simple, we introduce a variational version of single-task learning (VSTL), and provide the derivation of its evidence lower bound. It is worth noting that single-task learning does not share knowledge among tasks. Thus, the log-likelihood for single-task learning is not conditioned on the data from other related tasks.

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{x}) &= \log \int\int p(\mathbf{y}, \mathbf{z}, \mathbf{w}, |\mathbf{x}) d\mathbf{w} d\mathbf{z} \\
&= \log \int\int p(\mathbf{y}|\mathbf{z}, \mathbf{w}) p(\mathbf{w}) p(\mathbf{z}|\mathbf{x}) d\mathbf{w} d\mathbf{z} \\
&= \log \int\int p(\mathbf{y}|\mathbf{z}, \mathbf{w}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z} p(\mathbf{w}) d\mathbf{w} \\
&= \log \int \frac{[\int p(\mathbf{y}|\mathbf{z}, \mathbf{w}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z}] p(\mathbf{w}) q(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} \\
&\geq -\mathbb{KL}(q(\mathbf{w})||p(\mathbf{w})) + \mathbb{E}_{q(\mathbf{w})}[\log \int p(\mathbf{y}|\mathbf{z}, \mathbf{w}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z}] \\
&\geq -\mathbb{KL}[q(\mathbf{w})||p(\mathbf{w})] + \mathbb{E}_{q(\mathbf{w})}[\log \int \frac{p(\mathbf{y}|\mathbf{z}, \mathbf{w}) p(\mathbf{z}|\mathbf{x}) q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})} d\mathbf{z}] \\
&\geq -\mathbb{KL}[q(\mathbf{w})||p(\mathbf{w})] - \mathbb{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] + \mathbb{E}_{q(\mathbf{w}), q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{y}|\mathbf{z}, \mathbf{w})]
\end{aligned} \tag{16}$$

Usually, the approximate posteriors $q(\mathbf{w})$ and $q(\mathbf{z}|\mathbf{x})$ are defined as a fully-factorized Gaussian distribution. Due to lack of extracted information offered by other related tasks, the priors $p(\mathbf{w})$ and $p(\mathbf{z}|\mathbf{x})$ are set to a standard Gaussian distribution, as applied in (Sohn et al., 2015; Kingma et al., 2015; Molchanov et al., 2017).

B EXTRA EXPERIMENTAL RESULTS

B.1 EFFECTIVENESS IN HANDLING LIMITED DATA

We further provide detailed information in Table 6 about average accuracy in Fig. 2. Our proposed probabilistic models, i.e., VMTL and VMTL-AC outperform the deterministic baseline multi-task learning model (bMTL), which demonstrates the benefits of our proposed variational Bayesian framework. Given a limited amount of training data, STL and VSTL can not train a proper model for each task. As the training data decreases, our methods based on the variational Bayesian framework are able to better handle this challenging case by incorporating the shared knowledge into the prior of each tasks. The best results of average accuracy are marked in bold, while the second best by underline.

Table 6. Performance of average accuracy under different proportions of training data on *Office-Home*.

Methods	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
STL	49.2	58.3	61.3	64.9	66.4	67.7	68.3	70.3	70.4	71.9
VSTL	51.1	60.2	63.0	65.8	67.9	69.6	70.3	**72.4**	**72.3**	**73.8**
bMTL	50.4	59.5	62.4	65.6	66.8	68.3	68.9	70.5	70.7	72.3
VMTL-AC	56.6	63.7	65.4	68.1	68.7	69.4	70.0	70.9	70.4	71.5
VMTL	**58.2**	**65.0**	**66.4**	**69.1**	**69.9**	**70.4**	**71.3**	72.2	72.0	73.3

B.2 EFFECTIVENESS VARIATIONAL BAYESIAN APPROXIMATION

The comparison results on performance of Bayesian approximation for representations $\mathbf{z}$ and classifiers $\mathbf{w}$ on the Office-Home, Office-Caltech and ImageCLEF datasets are shown in Tables 7, 8 and 9, respectively. Both variational Bayesian representations and classifiers can benefit the performance. And we find that Bayesian classifiers in the variational inference framework contribute more to the performance than Bayesian representations. It is likely due to the fact that Bayesian classifiers can better improve the model's discriminative ability. Our method jointly infers the posteriors over feature representations and classifiers in a Bayesian framework, which consistently outperforms its variants on three benchmarks.

Table 7. Detailed results on performance of Bayesian approximation for representation $\mathbf{z}$ and classifier $\mathbf{w}$ on *Office-Home*.

$\mathbf{z}$	$\mathbf{w}$	5%					10%					20%				
		A	C	P	R	Avg.	A	C	P	R	Avg.	A	C	P	R	Avg.
×	×	37.6	31.5	68.5	63.8	50.4	51.0	41.6	76.0	69.2	59.5	56.6	51.8	80.9	72.9	65.6
✓	×	39.2	32.7	69.1	64.4	51.4	52.7	42.5	76.6	70.2	60.5	57.8	53.2	81.8	74.1	66.7
×	✓	51.6	37.8	70.7	66.9	56.8	57.6	47.2	77.4	72.5	63.7	62.5	53.5	82.0	76.0	68.5
✓	✓	53.8	38.6	71.4	68.8	**58.2**	60.3	47.5	78.1	74.2	**65.0**	64.0	53.3	82.5	76.6	**69.1**

Table 8. Detailed results on performance of Bayesian approximation for representation $\mathbf{z}$ and classifier $\mathbf{w}$ on *Office-Caltech*.

$\mathbf{z}$	$\mathbf{w}$	5%					10%					20%				
		A	W	D	C	Avg.	A	W	D	C	Avg.	A	W	D	C	Avg.
×	×	90.0	89.4	95.0	83.5	89.5	93.6	97.0	92.1	86.3	92.3	95.0	94.5	96.0	86.8	93.1
✓	×	91.0	91.3	97.9	84.5	91.2	94.4	97	94.3	87.1	93.2	95.7	94.9	95.2	87.5	93.3
×	✓	93.2	95.5	95.7	89.6	93.5	94.8	97.3	97.8	90.6	95.1	95.6	96.6	99.2	89.8	**95.3**
✓	✓	93.8	95.5	96.4	90.0	**93.9**	95.5	97.0	97.9	91.0	**95.3**	95.6	95.8	99.2	90.6	**95.3**

B.3 EFFECTIVENESS OF GUMBEL-SOFTMAX PRIORS

The performance comparison of the proposed VMTL with different priors on the Office-Home, Office-Caltech and ImageCLEF datasets is shown in Tables 10, 11 and 12, respectively. "Mean" denotes that the prior of current takes is the mean of variational posteriors of other related tasks. "Learnable weighted" denotes that weights of mixing the variational posteriors of other related tasks

Table 9. Detailed results on performance of Bayesian approximation for representation **z** and classifier **w** on *ImageCLEF*.

z	**w**	5%					10%					20%				
		C	I	P	B	Avg.	C	I	P	B	Avg.	C	I	P	B	Avg.
×	×	88.3	73.2	61.2	40.0	65.7	90.6	79.3	66.3	51.9	72.0	92.9	86.5	71.9	56.0	76.8
✓	×	89.0	73.5	62.6	42.4	66.9	91.3	79.8	68	51.5	72.6	93.5	88.1	74.0	57.5	78.3
×	✓	91.4	81.9	71.8	58.4	75.9	93.0	86.3	72.0	60.0	77.8	93.8	88.1	77.1	59.0	79.5
✓	✓	91.1	83.2	71.4	58.3	**76.0**	93.7	86.5	71.8	59.5	**77.9**	94.0	89.7	77.9	59.7	**80.3**

are learnable. Our Gumbel-softmax Priors apply Gumbel-softmax techinique to learn the mixing weights, which introduces uncertainty to the relationships among tasks in order to explore sufficient transferable information from other tasks. In the three datasets, our designed priors outperform other methods consistently.

Table 10. Detailed results on performance of VMTL with different priors on *Office-Home*.

Priors	5%					10%					20%				
	A	C	P	R	Avg.	A	C	P	R	Avg.	A	C	P	R	Avg.
Mean	52.0	38.0	70.6	68.0	57.2	59.7	47.0	77.9	73.5	64.5	63.0	53.0	82.0	76.5	68.6
Learnable weighted	51.4	38.0	70.7	67.7	57.0	59.8	47.2	77.8	72.7	64.4	63.1	53.3	82.2	75.9	68.6
Gumbel-Softmax	53.8	38.6	71.4	68.8	**58.2**	60.3	47.5	78.1	74.2	**65.0**	64.0	53.3	82.5	76.6	**69.1**

Table 11. Detailed results on performance of VMTL with different priors on *Office-Caltech*.

Priors	5%					10%					20%				
	A	W	D	C	Avg.	A	W	D	C	Avg.	A	W	D	C	Avg.
Mean	93.8	95.8	97.1	88.5	93.8	95.3	97.7	97.1	90.4	95.1	95.7	96.2	98.4	90.0	95.1
Learnable weighted	92.9	94.7	96.4	89.9	93.5	95.5	97.0	97.9	90.2	95.2	95.8	94.5	99.2	90.7	95.1
Gumbel-softmax	93.8	95.5	96.4	90.0	**93.9**	95.5	97.0	97.9	91.0	**95.3**	95.6	95.8	99.2	90.6	**95.3**

B.4 A COMPREHENSIVE COMPARISON WITH OTHER METHODS

The comprehensive comparison with state-of-the-art methods, including multi-task feature learning (MTFL) (Argyriou et al., 2007), robust multi-task learning (RMTL) (Chen et al., 2011), multi-task relationship learning (MTRL) Zhang & Yeung (2012), deep multi-task learning with tensor factorization (DMRL-TF) (Yang & Hospedales, 2016) and multilinear relationship network (MRN) (Long et al., 2017) is shown in Table 13. The results of the above state-of-the-art methods are taken from paper (Long et al., 2017). The proposed VMTL consistently achieves the best performance on all datasets with all train-test split settings. VMTL with amortized classifiers (VMTL-AC) can produce competive performance better than most of previouss methods.

Table 12. Detailed results on performance of VMTL with different priors on *ImageCLEF*.

Priors	5%					10%					20%				
	C	I	P	B	Avg.	C	I	P	B	Avg.	C	I	P	B	Avg.
Mean	89.5	83.0	71.4	58.8	75.7	93.5	86.5	71.7	60.2	78.0	93.1	89.6	78.8	58.8	80.1
Learnable weighted	90.7	81.8	71.9	57.4	75.5	93.7	86.1	71.9	59.3	77.8	93.3	89.6	77.7	59.2	79.9
Gumbel-softmax	91.1	83.2	71.4	58.3	**76.0**	93.7	86.5	71.8	59.5	**77.9**	94.0	89.7	77.9	59.7	**80.3**

Table 13. Performance comparison of different methods on the *Office-Home* dataset for multiple tasks: Artistic (A), Clipart (C), Product (P) and Real-world (R).

Methods	**5%**					**10%**					**20%**				
	A	C	P	R	Avg.	A	C	P	R	Avg.	A	C	P	R	Avg.
MTFL	40.1	30.4	61.5	59.5	47.9	50.3	35.0	66.3	65.0	54.2	55.2	38.8	69.1	70.0	58.3
RMTL	42.3	32.8	62.3	60.6	49.5	49.7	34.6	65.9	64.6	53.7	55.2	39.2	69.6	70.5	58.6
MTRL	42.7	33.3	62.9	61.3	50.1	51.6	36.3	67.7	66.3	55.5	55.8	39.9	70.2	71.2	59.3
DMTL-TF	49.2	34.5	67.1	62.9	53.4	57.2	42.3	73.6	69.9	60.8	58.3	56.1	79.3	72.1	66.5
MRN	53.3	36.4	70.5	67.7	57.0	59.9	42.7	76.3	73.0	63.0	58.5	55.6	80.7	72.8	66.9
VMTL-AC	52.3	37.5	70.1	66.7	56.7	58.4	46.5	76.9	73.1	63.7	62.3	52.4	82.0	75.9	68.2
VMTL	53.8	38.6	71.4	68.8	**58.2**	60.3	47.5	78.1	74.2	**65.0**	64.0	53.3	82.5	76.6	**69.1**