Multi-Task Learning: Unlocking Potential with New Python Library

Kirill Semlin Iryna Zabarianska Ilgam Latypov
Alexander Terentyev

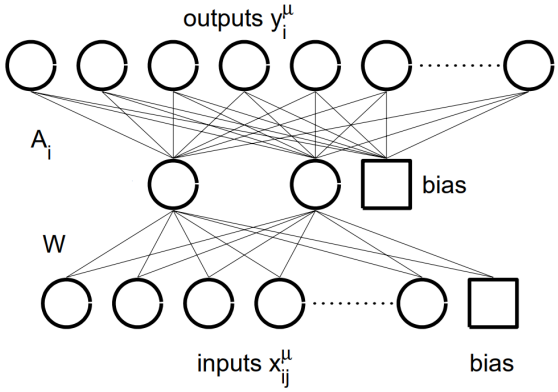
Multi-task learning (MTL) has emerged as a powerful paradigm in machine learning, aiming to leverage the relationships among related tasks to enhance the performance of individual tasks. This approach is particularly beneficial in scenarios where data is scarce for each task, making it challenging to develop robust models independently. By recognizing that many real-world problems can be framed as a series of similar yet distinct tasks, MTL provides a framework for sharing knowledge across these tasks, ultimately leading to improved outcomes.

The central objective of MTL is to explore and exploit task relatedness to bolster the performance of each individual task. However, uncovering these relationships can be complex and highly nonlinear, presenting significant challenges.

This article presents an overview of various models for MTL and announces a new Python library, <**name of lib**>, featuring their implementation.

1 Task Clustering and Gating for MTL

Suppose that for task i we are given a data set $D_i = \{x_i^\mu, y_i^\mu\}$, where μ the number of examples for task i. Our focus is on a neural network architecture where the input-to-hidden weights W are shared across all tasks, while each task maintains its own set of task-specific output weights A_i:



The choice of prior for the output weights A_i significantly influences the resulting algorithms we can derive from this framework. More details in the article [1].

Clustering of Tasks

Suppose we have several clusters of similar tasks. We could take as a prior a mixture of $n_{cluster}$ Gaussians

$$A_i \sim \sum_{\alpha=1}^{n_{cluster}} q_\alpha \mathcal{N}(m_\alpha, \Sigma_\alpha),$$

where q_α represents the *a priori* probability for any task to be 'assigned to' cluster α. Then the posterior data likelihood expressed as

$$P(D_i|\Lambda) = \int dA_i P(D_i|A_i, \Lambda) \sum_{\alpha=1}^{n_{cluster}} q_\alpha P(A_i|m_\alpha, \Sigma_\alpha).$$

In this way, the posterior distribution effectively 'assigns' tasks to that cluster that is most compatible with the data within the task, in the sense that all other clusters (Gaussians) do contribute much less.

Gating of Tasks

In task clustering approach the prior is task-independent: all tasks are assigned to each of the clusters with the same probabilities q_α. A natural extension is to incorporate the task-dependent features f_i in a gating model, for example

$$q_{i\alpha} = e^{U_\alpha^T f_i} / \sum_{\alpha'} e^{U_{\alpha'}^T f_i}.$$

The above task clustering approach is a special case with $n_{feature} = 1$ and $f_i = 1$ for all tasks i.

2 MTL with Latent Hierarchiess

Our model for Domain Adaptation and Multi-Task Learning leverages a latent hierarchical structure to capture task relationships. Specifically, we adopt Kingman's N-coalescent as a nonparametric Bayesian prior, which provides a flexible framework for modeling dependencies between tasks without predefined constraints. More details in the article [2].

Domain Adaptation

A tree structure will be generated according to a K-coalescent process, and weight vectors will be propagated along this tree. The root of the tree represents the "global" weight vector, while the leaves correspond to the weight vectors specific to each task. It is assumed that the weight vectors evolve according to Brownian diffusion.

Multi-Task Learning

The algorithm for MTL will be a modification of the algorithm used for DA. In the case of MTL, weight vectors will not be shared; instead, their covariance structure will be shared. This model is somewhat more complex to specify because Brownian motion is not applicable to a covariance structure (for instance, it does not preserve positive semi-definiteness). To address this issue, the covariance structure will be decomposed into correlations and standard deviations, assuming a constant global correlation matrix while allowing the standard deviations to vary across the tree.

3 Sparse Bayesian MTL

One of the challenges in multi-task learning involves effectively managing the covariance structure of weight matrices W to enable simultaneous learning of correlated and anti-correlated tasks. Traditional models often rely on a block diagonal covariance structure, which restricts their capacity to capture complex relationships between tasks.

To address this issue, a hierarchical Bayesian approach is employed that introduces sparsity into the weight matrix W. This method utilizes a matrix-variate Gaussian scale mixture as an effective prior for each block W_i. This approach facilitates the reduction of entire blocks of weights to zero, promoting sparsity while ensuring computational feasibility. For further details, refer to the article [3].

A general family of group sparsity inducing priors

For further construction of sparse models and variational inference, the method used to construct the general family of priors that induce group sparsity is of critical significance.

We will assume that the effective prior on each block $W_i \in \mathbb{R}^{P \times D_i}$ has the form of a matrix-variate Gaussian scale mixture, extending the multivariate Gaussian scale mixture:

$$p(W_i) = \int_0^\infty \mathcal{N}(0, \gamma_i^{-1}\Omega_i, \Sigma) p(\gamma_i) d\gamma_i,$$

where Ω_i, Σ and γ_i is the latent precision for block W_i.

To induce sparsity in W_i, we select an appropriate hyperprior for γ_i. Then impose a generalized inverse Gaussian prior for the latent precision variables:

$$\gamma_i \sim \mathcal{N}^{-1}(\omega, \chi, \phi) = \frac{\chi^{-\omega}(\sqrt{\chi\phi})^\omega}{2K_\omega\sqrt{\chi\phi}} \gamma_i^{\omega-1} e^{-\frac{1}{2}(\chi\gamma_i^{-1} + \phi\gamma_i)}$$

where $K_\omega(.)$ is the modified Bessel function of the second kind, ω is the index, $\sqrt{\chi\phi}$ defines the concentration of the distribution and $\sqrt{\chi/\phi}$ defines its scale.

Consequently, the effective prior becomes a symmetric matrix-variate generalized hyperbolic distribution:

$$p(W_i) \propto \frac{K_{\omega+\frac{PD_i}{2}}\left(\sqrt{\chi(\phi + tr\{\Omega_i^{-1}W_i^T\Sigma^{-1}W_i\}}\right)}{\left(\sqrt{\frac{\chi(\phi+tr\{\Omega_i^{-1}W_i^T\Sigma^{-1}W_i\}}{\chi}}\right)^{\omega+\frac{PD_i}{2}}}.$$

4 Variational MTL with Gumbel-Softmax Priors

Another scenario of multi-task learning occurs when the tasks share the same label or target space but are characterized by different data distributions. Each task, facing the challenge of limited labeled data, benefits from a probabilistic framework that captures uncertainty and enhances model performance.

Variational MTL approach specifies the prior for each task's classifier based on variational posteriors from other tasks, allowing for a principled exploration of inter-task relationships. To optimize knowledge sharing while minimizing interference from irrelevant tasks, the Gumbel-Softmax technique is implemented. More details in the article [4].

Learning Task Relatedness via Gumbel-Softmax Priors

To leverage the shared knowledge, we propose to specify the prior of the classifier for the current task t using the variational posteriors over classifiers of other tasks:

$$p(w_t^{(\eta)}|D_{1:T\setminus t}) = \sum_{i\neq t} \alpha_{ti} q_\theta(w_i^{(\eta-1)}|D_i),$$

where η is the iteration number, D_t training data, $w_t^{(\eta)}$ - latent variables.

During training, the goal is for each task to learn relevant shared knowledge from closely related tasks while minimizing interference from irrelevant ones. To achieve this, we employ the Gumbel-Softmax technique to learn mixing weights, defined as:

$$\alpha_{ti} = \frac{exp((log\pi_{ti} + g_{ti})/\tau}{\sum_{i\neq t} exp((log\pi_{ti} + g_{ti})/\tau}.$$

Here α_{ti} is the mixing weight that indicates the relatedness between tasks t and i. The parameter π_{ti} is a learnable component indicating the likelihood of knowledge transfer between tasks, while g_{ti} is sampled from a Gumbel distribution, using inverse transform sampling. The temperature parameter τ controls the softmax behavior. This approach helps manage potential negative transfer by encouraging lower mixing weights for less relevant tasks, effectively reducing interference.

Then we will define variational posteriors as fully factorized Gaussians for each class:

$$q_\theta(w_t|D_t) = \prod_{c=1}^{C} q_\theta(w_{t,c}|D_{t,c}) = \prod_{c=1}^{C} \mathcal{N}(\mu_{t,c}, diag(\sigma_{t,c}^2)).$$

5 New Python Library: <name of lib>

The models for MTL described above have been implemented in library <**name of lib**>.

We aim to deliver dependable, efficient, and user-friendly tools that will be advantageous for both researchers and practitioners. We anticipate that the library's implementation will foster further advancements in this scientific field. Keep an eye out for updates!

Git: https://github.com/intsystems/bmm-multitask-learning.git

References

[1] Bart Bakker and Tom Heskes, "Task Clustering and Gating for Bayesian Multitask Learning," Journal of Machine Learning Research, vol. 4, pp. 83–99, 2003.

[2] Hal Daum´e III, "Bayesian Multitask Learning with Latent Hierarchies," arXiv preprint arXiv preprint arXiv:2002.04799, 2020.

[3] Cedric Archambeau, Shengbo Guo and Onno Zoeter, "Sparse Bayesian Multi-Task Learning," Advances in Neural Information Processing Systems, vol. 24, 2011.

[4] Jiayi Shen, Xiantong Zhen, Marcel Worringa and Ling Shao, "Variational Multi-Task Learning with Gumbel-Softmax Priors," 35th Conference on Neural Information Processing Systems, 2021.