Многозадачное обучение

Московский Физико-Технический Институт

2022

Многозадачное обучение

Wiki

Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better.

Task Clustering and Gating for Bayesian Multitask Learning, 2003

Рассматривается нейросеть с общим скрытым слоем:

$$y^i = (\mathrm{x}\mathrm{W}_{\mathsf{shared}})\,\mathrm{W}^i_{\mathsf{task}}$$

Как определить взаимосвязь задач?

Task Clustering and Gating for Bayesian Multitask Learning, 2003

Как определить взаимосвязь задач?

- Никак: $\mathrm{W}^i_{\mathrm{task}} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$
- Через гауссову смесь: $\mathrm{W}^i_{\mathrm{task}} \sim \sum \alpha_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$
- Через gating-функцию: $\mathrm{W}^i_{\mathrm{task}} \sim \sum \mathcal{N}(\mathrm{softmax}(\boldsymbol{\mu}), \boldsymbol{\Sigma})$

Gating vs Mixutre

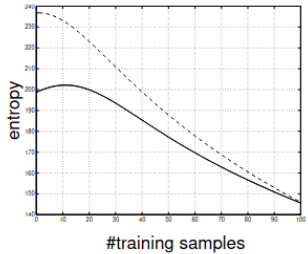


Figure 9: The entropy of the task-cluster assignment probabilities $\rho_{i\alpha}$ as a function of the number of samples for the task clustering method (dashed line) and the gating method (solid line). For few samples, the gating method clusters more strongly (has a lower entropy) than the task clustering method. For higher numbers of samples, the two methods behave similarly.

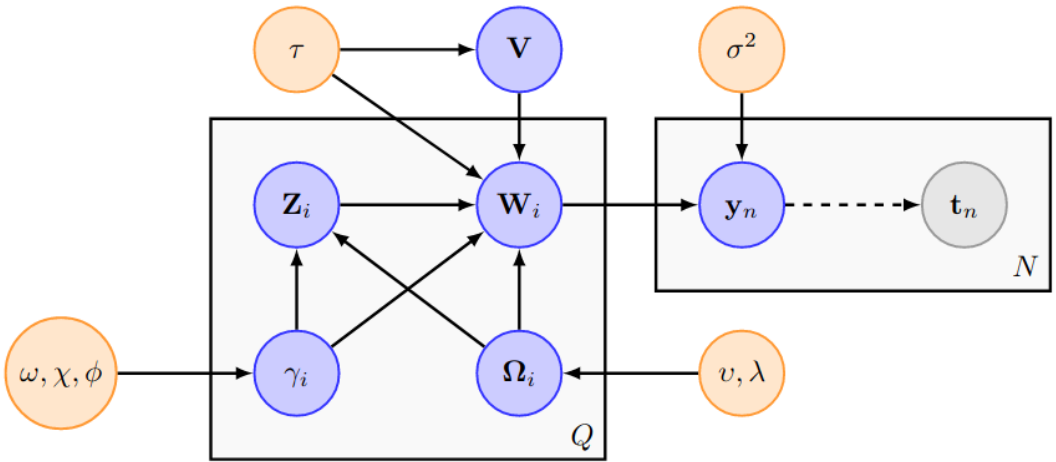
Многозадачное обучение и domain adaptation

Есть ли связь между многозадачным обучением и domain adaptation?

Sparse Bayesian Multi-Task Learning, 2011

$$\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n, \qquad \boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

Sparse Bayesian Multi-Task Learning, 2011

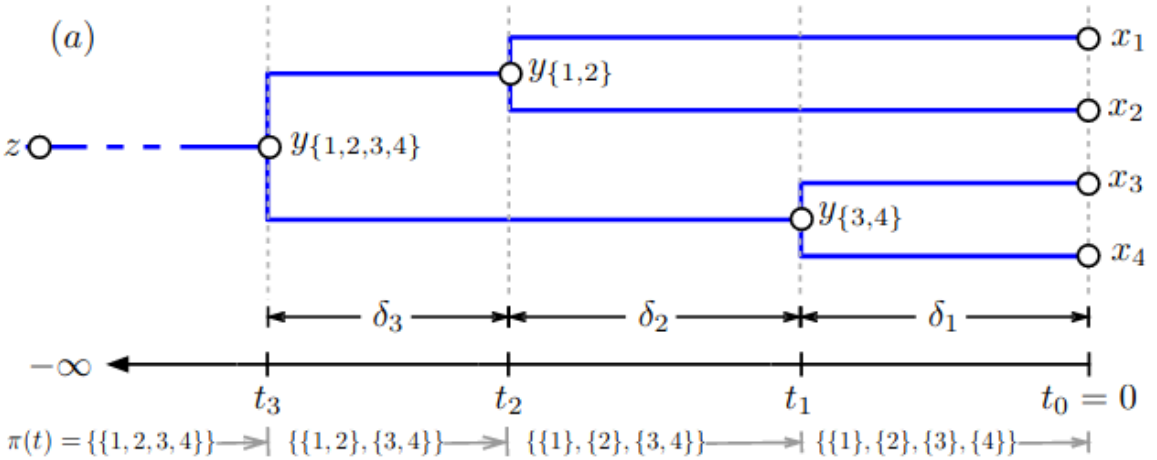


Bayesian Multitask Learning with Latent Hierarchies

Daume III, 2009

We exploit the intuition that for domain adaptation, we wish to share classifier structure, but for multitask learning, we wish to share covariance structure.

Bayesian Multitask Learning with Latent Hierarchies



Bayesian Multitask Learning with Latent Hierarchies

1. Choose a global *mean* and *covariance* $(\boldsymbol{\mu}^{(0)}, \boldsymbol{\Lambda}) \sim \mathcal{N}or\mathcal{IW}(0, \sigma^2\mathbf{I}, D+1)$. [2]
2. Choose a tree structure $(\pi, \boldsymbol{\delta}) \sim$ *Coalescent* over K leaves.
3. For each non-root node i in π (top-down):
   (a) Choose $\boldsymbol{\mu}^{(i)} \sim \mathcal{N}or(\boldsymbol{\mu}^{(p_\pi(i))}, \delta_i\boldsymbol{\Lambda})$, where $p_\pi(i)$ is the parent of i in π.
4. For each domain $k \in [K]$:
   (a) Denote by $\boldsymbol{w}^{(k)} = \boldsymbol{\mu}^{(i)}$ where i is the leaf in π corresponding to domain k.
   (b) For each example $n \in [N_k]$:
      i. Choose input $\boldsymbol{x}_n^{(k)} \sim \mathcal{D}^{(k)}$.
      ii. Choose output $y_n^{(k)}$ by:
      Regression: $\mathcal{N}or(\boldsymbol{w}^{(k)\top}\boldsymbol{x}_n^{(k)}, \rho^2)$
      Classification: $\mathcal{B}in(1/(1+e^{-\boldsymbol{w}^{(k)\top}\boldsymbol{x}_n^{(k)}}))$

Bayesian Multitask Learning with Latent Hierarchies

→ 1. Choose $\mathbf{R}$ by Eq (2) and deviation covariance $\mathbf{\Lambda} \sim \mathcal{IW}(\sigma^2\mathbf{I}, D+1)$.

2. Choose a tree structure $(\pi, \boldsymbol{\delta}) \sim$ *Coalescent* over K leaves.

3. For each non-root node i in π (top-down):

→ (a) Choose $\mathbf{S}^{(i)} \sim \mathcal{N}or(\mathbf{S}^{(p_\pi(i))}, \delta_i \mathbf{\Lambda})$, where $p_\pi(i)$ is the parent of i in π.

4. For each task $k \in [K]$:

→ (a) Choose $\boldsymbol{w}^{(k)}$ by (i is the leaf associated with task k): $\mathcal{N}or\big(0, \big(\exp \mathbf{S}^{(i)}\big)\mathbf{R}\big(\exp \mathbf{S}^{(i)}\big)\big)$

(b) For each example $n \in [N_k]$:

→ i. Choose input $\boldsymbol{x}_n^{(k)} \sim \mathcal{D}$.

ii. Choose output $y_n^{(k)}$ by:

Regression: $\mathcal{N}or(\boldsymbol{w}^{(k)\top}\boldsymbol{x}_n^{(k)}, \rho^2)$

Classification: $\mathcal{B}in(1/(1+e^{-\boldsymbol{w}^{(k)\top}\boldsymbol{x}_n^{(k)}}))$

Automated Curriculum Learning, 2017

Algorithm 1 Intrinsically Motivated Curriculum Learning

Initially: $w_i = 0$ for $i \in [N]$

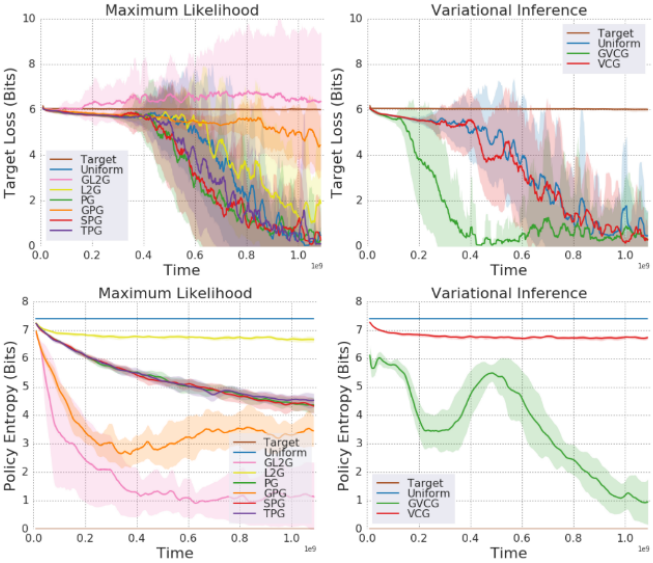
for $t = 1 \dots T$ **do**
- $\pi(k) := (1-\epsilon)\frac{e^{w_k}}{\sum_i e^{w_i}} + \frac{\epsilon}{N}$
- Draw task index k from π
- Draw training sample $\mathbf{x}$ from D_k
- Train network p_θ on $\mathbf{x}$
- Compute learning progress ν (Sections 3.1 & 3.2)
- Map $\hat{r} = \nu/\tau(\mathbf{x})$ to $r \in [-1, 1]$ (Section 2.3)
- Update w_i with reward r using Exp3.S (1)

end for

Automated Curriculum Learning, 2017

- Loss-driven Progress
  - Prediction Gain: $L(\mathrm{w}', \mathrm{x}) - L(\mathrm{w}, \mathrm{x})$
  - Gradient prediction gain
  - Self prediction gain: сэмплируем x
  - Mean prediction gain: усредняем по всем задачам.
- Complexity-driven Progress
  - Variational complexity gain: $KL(q'|p) - KL(q|p)$
  - Gradient Variational complexity gain
  - L2G: разница в l2-регуляризации
  - GL2G: градиент по L2G

Automated Curriculum Learning, 2017



Continual learning

Continual learning

Continual Learning is a concept to learn a model for a large number of tasks sequentially without forgetting knowledge obtained from the preceding tasks, where the data in the old tasks are not available any more during training new ones.

Three scenarios for continual learning

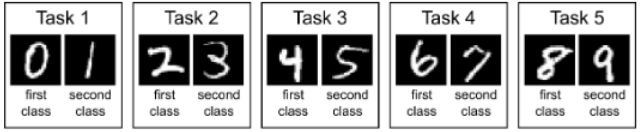
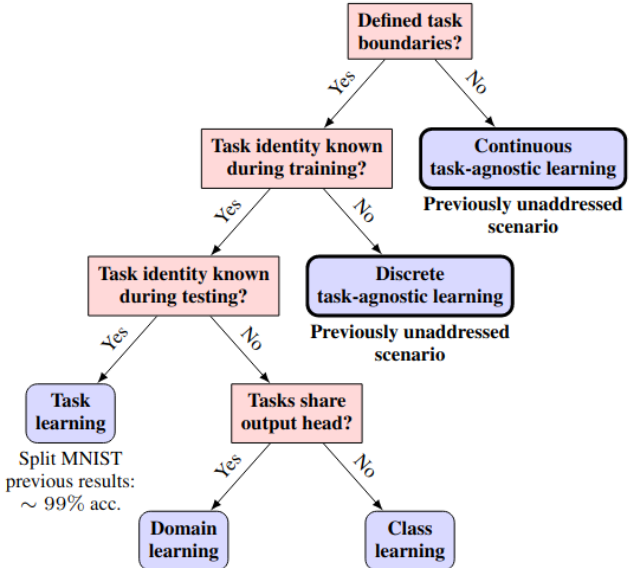


Figure 1: Schematic of split MNIST task protocol.

Table 2: Split MNIST according to each scenario.

Task-IL	With task given, is it the 1st or 2nd class? (e.g., 0 or 1)
Domain-IL	With task unknown, is it a 1st or 2nd class? (e.g., in $[0, 2, 4, 6, 8]$ or in $[1, 3, 5, 7, 9]$)
Class-IL	With task unknown, which digit is it? (i.e., choice from 0 to 9)

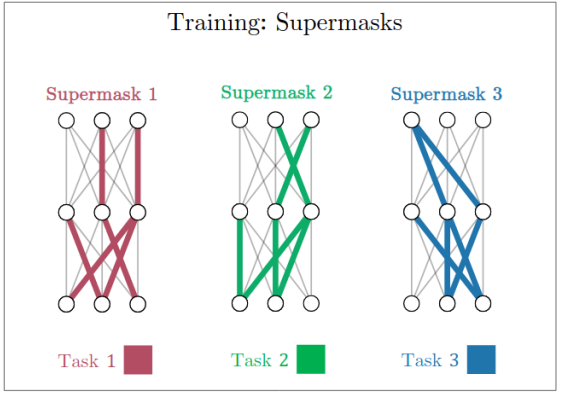
Task Agnostic Continual Learning Using Online Variational Bayes

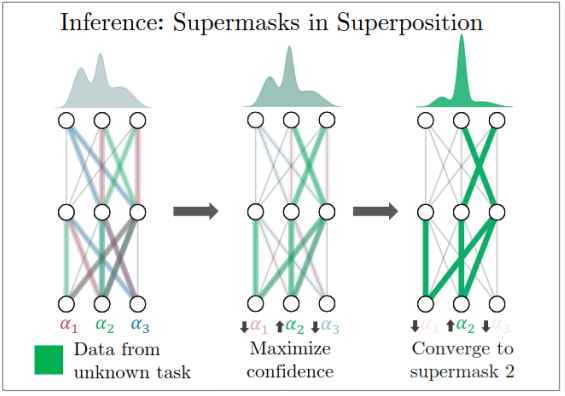


Continual learning: Типизация задач

Scenario	Description	Task space discreet or continuous?	Example methods / task names used
GG	Task **G**iven during train and **G**iven during inference	Either	PNN [42], BatchE [51], PSP [4], “Task learning” [55], “Task-IL” [49]
GNs	Task **G**iven during train, **N**ot inference; **s**hared labels	Either	EWC [23], SI [54], “Domain learning” [55], “Domain-IL” [49]
GNu	Task **G**iven during train, **N**ot inference; **u**nshared labels	Discrete only	“Class learning” [55], “Class-IL” [49]
NNs	Task **N**ot given during train **N**or inference; **s**hared labels	Either	BGD, “Continuous/discrete task agnostic learning” [55]

Supermasks in Superposition





Литература

- Bakker B. J., Heskes T. M. Task clustering and gating for bayesian multitask learning. – 2003.
- Guo S., Zoeter O., Archambeau C. Sparse Bayesian multi-task learning //Advances in Neural Information Processing Systems. – 2011. – T. 24.
- Daumé III H. Bayesian multitask learning with latent hierarchies //Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. – 2009. – C. 135-142.
- Graves A. et al. Automated curriculum learning for neural networks //international conference on machine learning. – PMLR, 2017. – C. 1311-1320.
- Van de Ven G. M., Tolias A. S. Three scenarios for continual learning //arXiv preprint arXiv:1904.07734. – 2019.
- Zeno C. et al. Task agnostic continual learning using online variational bayes //arXiv preprint arXiv:1803.10123. – 2018.
- Wortsman M. et al. Supermasks in superposition //Advances in Neural Information Processing Systems. – 2020. – T. 33. – C. 15173-15184.