uai2022

Exploiting Inferential Structure in Neural Processes

Dharmesh Tailor[1], Emtiyaz Khan[2], Eric Nalisnick[1]

[1]University of Amsterdam (Netherlands) [2]RIKEN AIP (Japan)

University of Amsterdam

RIKEN AIP

Main Takeaway

- **Problem:** Neural processes (NPs) perform fast-adaptation to new tasks at test-time but their training procedure is data-inefficient requiring a wide range of datasets to generalize well.
- **Main contribution:** We propose to incorporate a structured-inference network (SIN) [1]
- **Technical contributions:**
  - i. priors can be naturally incorporated
  - ii. leads to aggregation strategies in which context points are appropriately weighted
  - iii. interpretability of datapoint-wise encodings as neural sufficient statistics

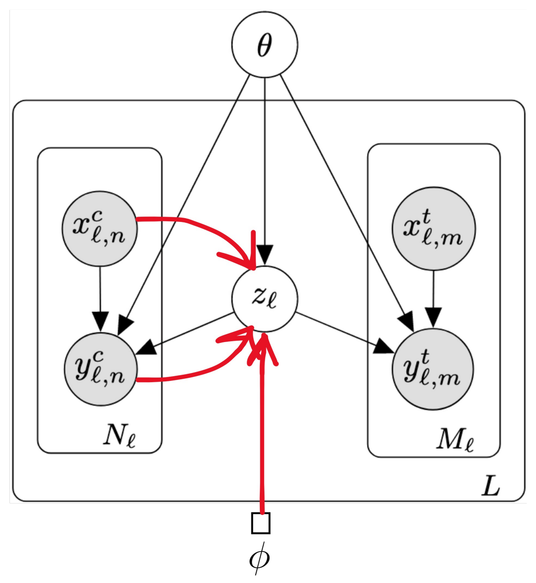
Background: Meta-Learning

- **Meta-learning**: NPs generalize between multiple, related tasks by modelling task-relatedness using hierarchical Bayes [2]
- **Train-time:**

$$\max \sum_{l=1}^{L} \log p_\theta\left(\mathcal{D}_l^t | \mathcal{D}_l^c\right)$$

$$= \sum_{l=1}^{L} \log \int p_\theta\left(\mathcal{D}_l^t | z_l\right) p_\theta\left(z_l | \mathcal{D}_l^c\right) dz_l$$

$$\gtrapprox \sum_{l=1}^{L} \mathbb{E}_{q_\phi(z|\mathcal{D}_l)}\left[\log p_\theta\left(\mathcal{D}_l^t | z\right)\right] - \mathrm{D}_{\mathrm{KL}}\left[q_\phi\left(z|\mathcal{D}_l^t \cup \mathcal{D}_l^c\right) \| q_\phi\left(z|\mathcal{D}_l^c\right)\right]$$



Amortised inference network

- **Test-time**

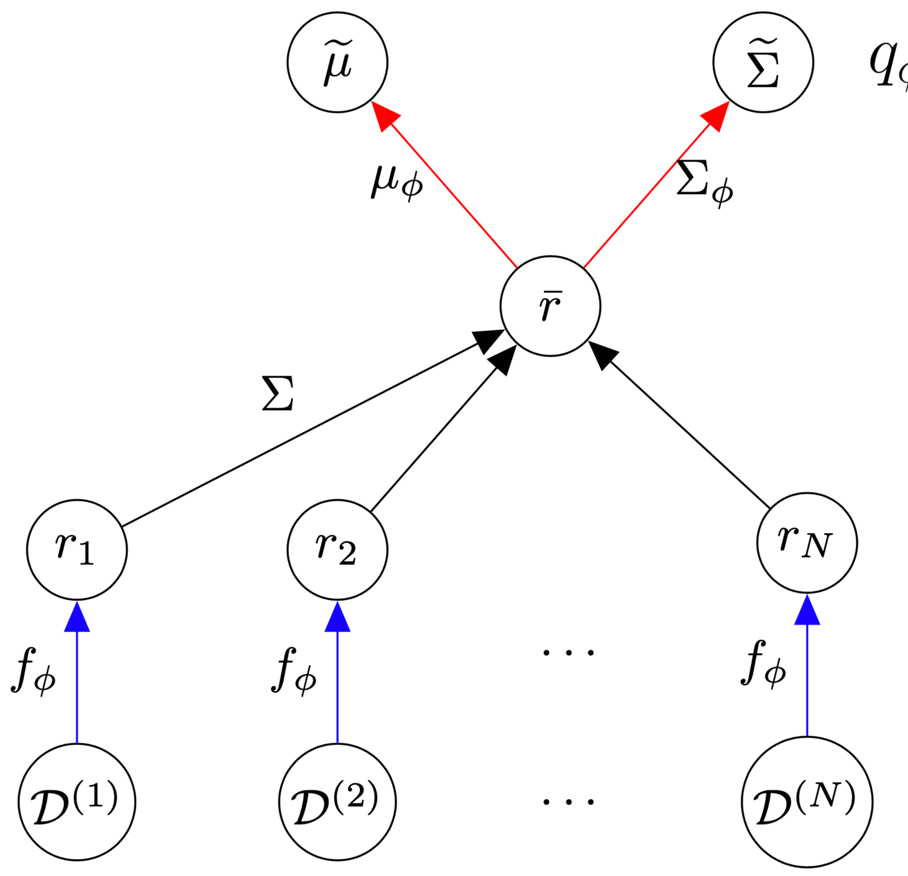
$$\mathbb{E}_{z \sim q_\phi(z|\mathcal{D}_*^c)}\left[\prod_{(x,y)\in\mathcal{D}_*^t} p_\theta(y|x,z)\right]$$

Background: Sum-Decomposition Network

Amortised inference network needs to:

- process datasets of variable-size
- permutation-invariant to the ordering of datapoints [3]

$$q_\phi(z|\mathcal{D}) = \mathbb{N}\left(\mu_\phi(\bar{\mathbf{r}}), \Sigma_\phi(\bar{\mathbf{r}})\right)$$



Aggregation:

$$\bar{\mathbf{r}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{r}_i$$

Datapoint-wise encodings:

$$\mathbf{r}_i = f_\phi(\mathcal{D}^{(i)})$$

Structured-Inference Network

Conjugate-exponential family

$$p(z|\mathcal{D}) = \frac{1}{p(\mathcal{D})}\prod_{i=1}^{N} p(y_i|z)\, p(z)$$

$$\propto h(z)\exp\left[\left\langle T(z), \sum_{i=1}^{N} \eta_i(y_i) + \eta_0 \right\rangle\right]$$

Sufficient statistics of likelihood

Natural parameters of prior

Combine recognition networks with conjugate-computations:

$$q_\phi(z|\mathcal{D}) = h(z)\exp\left[\left\langle T(z), \sum_{i=1}^{N} f_\phi(\mathcal{D}^{(i)}) + \eta_0 \right\rangle\right]$$

Neural Sufficient Statistics

Bayesian Aggregation

Given an exp-family prior, aggregation strategy naturally arises from the choice of parameterisation

For example, Gaussian prior & moment parameterisation recovers a recently proposed weighted aggregation strategy [4]

$$\{\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\} \leftrightarrow \eta_0$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{N} \mathbf{V}_i^{-1} + \boldsymbol{\Sigma}_0^{-1}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\sum_{i=1}^{N} \mathbf{V}_i^{-1}\mathbf{m}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

$$\mathbf{m}_i, \mathbf{V}_i = f_\phi(\mathcal{D}^{(i)})$$

Extension to structured priors such as mixture of Gaussian and Student's T via minimal conditional-EF form [5]

References

1. Lin, W., et al. Variational Message Passing with Structured Inference Networks. International Conference on Learning Representations, 2018.
2. Heskes, T. Empirical Bayes for Learning to Learn. International Conference on Machine Learning, 2000.
3. Zaheer, M., et al. Deep Sets. Neural information processing systems, 2017.
4. Volpp, M., et al. Bayesian Context Aggregation for Neural Processes. International Conference on Learning Representations, 2020.
5. Lin, W., et al. Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-Family Approximations. International Conference on Machine Learning, 2019.

OpenReview.net