Bayesian Context Aggregation for Neural Processes

Michael Volpp^{1,2}, Fabian Flürenbrock¹, Lukas Grossberger¹, Christian Daniel¹, Gerhard Neumann²

¹Bosch Center for Artificial Intelligence, Renningen, Germany ²Karlsruhe Institute of Technology, Karlsruhe, Germany



Probabilistic Regression as a Multi-task Learning Problem

• Family \mathcal{F} of functions $f_{\ell}: \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ ("tasks") with some form of shared structure



ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes B Robert Book Carbol 2021 All Lights reported also reparties an discosel evolution production of the distribution as well as in the event of andirations for industrial generative rights.



Probabilistic Regression as a Multi-task Learning Problem

- Family \mathcal{F} of functions $f_{\ell}: \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ ("tasks") with some form of shared structure
- Noisy evaluations $\mathcal{D}_{\ell} = \{(x_{\ell,i}, y_{\ell,i})\}_i$ with $y_{\ell,i} = f_{\ell}(x_{\ell,i}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$



ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes 6 Robert Rept. (appl. 40) 41 (dots reported also response and discoal englishing or englishing distribution as well as in the event of andirations for industrial generative risks.



Probabilistic Regression as a Multi-task Learning Problem

- Family \mathcal{F} of functions $f_{\ell}: \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ ("tasks") with some form of shared structure
- Noisy evaluations $\mathcal{D}_{\ell} = \{(x_{\ell,i}, y_{\ell,i})\}_i$ with $y_{\ell,i} = f_{\ell}(x_{\ell,i}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$
- Learn the predictive distribution p(y^t_ℓ | x^t_ℓ, D^c_ℓ) over target outputs, conditioned on a context set D^c_ℓ ⊂ D_ℓ





Neural network (NN)-based multi-task learning architecture



¹Garnelo et al., Conditional Neural Processes, ICML 2018, Garnelo et al. Neural Processes. arxiv/1807.01622 2018 ICLR 2021 J. M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, *Bayesian Context Aggregation for Neural Processes* 6 Netre Block 1942021. Hight envend, alto particle and ydogal, ediablatic, mediata in the overel adjulction for waland ingentry right.



Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|\mathcal{D}^c)$ of the target function from $\mathcal{D}^c = \{(x_n^c, y_n^c)\}_{n=1}^N$

¹Garnelo et al., Conditional Neural Processes, ICML 2018, Garnelo et al. Neural Processes. arxiv/1807.01622 2018
ICLR 2021 [M. Volpp, F. Flörenbrock, L. Grossberger, C. Daniel, G. Neumann, *Bayesian Context Aggregation for Neural Processes*⁶ Robert Boht Cent 2021. If intervent able regardle and quadra deplated interventation. edited in the overol adjustant for balanti property injunction.



Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|\mathcal{D}^c)$ of the target function from $\mathcal{D}^c = \{(x_n^c, y_n^c)\}_{n=1}^N$ a. Map each context tuple (x_n^c, y_n^c) onto a latent observation $r_n = \operatorname{enc}_1(x_n^c, y_n^c)$

¹Garnelo et al., Conditional Neural Processes, ICML 2018, Garnelo et al. Neural Processes. arxiv/1807.01622 2018
ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, *Bayesian Context Aggregation for Neural Processes*6 Robert Boch 2012. Might served als aggregation graduate adjustice modulation graduate approximation.



Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|\mathcal{D}^c)$ of the target function from $\mathcal{D}^c = \{(x_n^c, y_n^c)\}_{n=1}^N$

- a. Map each context tuple (x_n^c, y_n^c) onto a latent observation $r_n = \text{enc}_1(x_n^c, y_n^c)$
- b. Form an aggregated latent observation \bar{r} using mean aggregation: $\bar{r} = \frac{1}{N} \sum_{n=1}^{N} r_n$

¹Garnelo et al., Conditional Neural Processes, ICML 2018, Garnelo et al. Neural Processes, arxiv/1807.01622 2018

ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes



Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|\mathcal{D}^c)$ of the target function from $\mathcal{D}^c = \{(x_n^c, y_n^c)\}_{n=1}^N$

- a. Map each context tuple (x_n^c, y_n^c) onto a latent observation $r_n = \text{enc}_1(x_n^c, y_n^c)$
- b. Form an aggregated latent observation \bar{r} using mean aggregation: $\bar{r} = \frac{1}{N} \sum_{n=1}^{N} r_n$
- c. Map \bar{r} onto the parameters of the latent distribution: $(\mu_r, \sigma_r^2) = \text{enc}_2(\bar{r})$

¹Garnelo et al., Conditional Neural Processes, ICML 2018, Garnelo et al. Neural Processes, arxiv/1807.01622 2018

ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes



Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|\mathcal{D}^c)$ of the target function from $\mathcal{D}^c = \{(x_n^c, y_n^c)\}_{n=1}^N$

- a. Map each context tuple (x_n^c, y_n^c) onto a latent observation $r_n = \text{enc}_1(x_n^c, y_n^c)$
- b. Form an aggregated latent observation \bar{r} using mean aggregation: $\bar{r} = \frac{1}{N} \sum_{n=1}^{N} r_n$
- c. Map \bar{r} onto the parameters of the latent distribution: $(\mu_z, \sigma_z^2) = \text{enc}_2(\bar{r})$

2. Map samples $z \sim p(z|\mathcal{D}^c)$ onto a Gaussian output distribution: $(\mu_v, \sigma_v^2) = \text{dec}(z, x^t)$

¹Garnelo et al., Conditional Neural Processes, ICML 2018, Garnelo et al, Neural Processes, arxiv/1807.01622 2018

ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes



 Different areas in the (x, y)-space can have different task ambiguity (TA)





- Different areas in the (x, y)-space can have different task ambiguity (TA)
- Context tuples (x_n^c, y_n^c) with
 - high TA should have little influence on $p(z|\mathcal{D}^c)$
 - low TA should have large influence on $p(z|\mathcal{D}^c)$





- Different areas in the (x, y)-space can have different task ambiguity (TA)
- Context tuples (x_n^c, y_n^c) with
 - high TA should have little influence on $p(z|\mathcal{D}^c)$
 - low TA should have large influence on $p(z|\mathcal{D}^c)$
- Mean aggregation assigns the same weight

to all context tuples: $\bar{r} = \frac{1}{N} \sum_{n=1}^{N} r_n$





- Different areas in the (x, y)-space can have different task ambiguity (TA)
- Context tuples (x_n^c, y_n^c) with
 - high TA should have little influence on $p(z|\mathcal{D}^c)$
 - low TA should have large influence on $p(z|\mathcal{D}^c)$
- Mean aggregation assigns the same weight

to all context tuples: $\bar{r} = \frac{1}{N} \sum_{n=1}^{N} r_n$



How to efficiently incorporate task ambiguity into NP parameter inference?



Context aggregation and parameter inference should be treated as one holistic mechanism! Directly aggregate the context data into the statistical description of z!



Context aggregation and parameter inference should be treated as one holistic mechanism! Directly aggregate the context data into the statistical description of z!

▶ NP with our Bayesian aggregation (BA):





Context aggregation and parameter inference should be treated as one holistic mechanism! Directly aggregate the context data into the statistical description of z!

▶ NP with our Bayesian aggregation (BA):



Compare: NP with traditional mean aggregation (MA):



ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes © Robert Bosch GmbH 2021. All rights reserved, also regarding any disposal, exploitation, reproduction, editing, distribution, zo well as in the event of applications for industrial property rights.



Context aggregation as Bayesian inference





- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$





- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n, \sigma_{r_n}^2) = \operatorname{enc}(x_n^c, y_n^c)$





- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n, \sigma_{r_n}^2) = \operatorname{enc}(x_n^c, y_n^c)$
- Latent posterior: $p(z | \{r_n\}) = \mathcal{N}(z | \mu_z, \text{diag}(\sigma_z^2))$



$$\sigma_{z}^{2} = \left[\sum_{n=1}^{N} (\sigma_{r_{n}}^{2})^{-1}\right]^{-1}, \quad \mu_{z} = \sum_{n=1}^{N} \frac{\sigma_{z}^{2}}{\sigma_{r_{n}}^{2}} r_{n}$$



- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n, \sigma_{r_n}^2) = \operatorname{enc}(x_n^c, y_n^c)$
- Latent posterior: $p(z | \{r_n\}) = \mathcal{N}(z | \mu_z, \text{diag}(\sigma_z^2))$



$$\sigma_z^2 = \left[\sum_{n=1}^N \left(\sigma_{r_n}^2\right)^{-1}\right]^{-1}, \quad \mu_z = \sum_{n=1}^N \frac{\sigma_z^2}{\sigma_{r_n}^2} r_n$$

(for each latent dim.)

• r_n enters μ_z with learned weight $\sigma_z^2/\sigma_{r_n}^2$



- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n, \sigma_{r_n}^2) = \operatorname{enc}(x_n^c, y_n^c)$
- Latent posterior: $p(z | \{r_n\}) = \mathcal{N}(z | \mu_z, \text{diag}(\sigma_z^2))$



$$\sigma_{z}^{2} = \left[\sum_{n=1}^{N} \left(\sigma_{r_{n}}^{2}\right)^{-1}\right]^{-1}, \quad \mu_{z} = \sum_{n=1}^{N} \frac{\sigma_{z}^{2}}{\sigma_{r_{n}}^{2}} r_{n}$$

- r_n enters μ_z with learned weight $\sigma_z^2/\sigma_{r_n}^2$
- Principled quantification of task ambiguity



- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n, \sigma_{r_n}^2) = \operatorname{enc}(x_n^c, y_n^c)$
- Latent posterior: $p(z | \{r_n\}) = \mathcal{N}(z | \mu_z, \text{diag}(\sigma_z^2))$



$$\sigma_{z}^{2} = \left[\sum_{n=1}^{N} \left(\sigma_{r_{n}}^{2}\right)^{-1}\right]^{-1}, \quad \mu_{z} = \sum_{n=1}^{N} \frac{\sigma_{z}^{2}}{\sigma_{r_{n}}^{2}} r_{n}$$

- r_n enters μ_z with learned weight $\sigma_z^2/\sigma_{r_n}^2$
- Principled quantification of task ambiguity
- Only marginal computational overhead



- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z, \operatorname{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n, \sigma_{r_n}^2) = \operatorname{enc}(x_n^c, y_n^c)$
- Latent posterior: $p(z | \{r_n\}) = \mathcal{N}(z | \mu_z, \text{diag}(\sigma_z^2))$



$$\sigma_z^2 = \left[\sum_{n=1}^N \left(\sigma_{r_n}^2\right)^{-1}\right]^{-1}, \quad \mu_z = \sum_{n=1}^N \frac{\sigma_z^2}{\sigma_{r_n}^2} r_n$$

- r_n enters μ_z with learned weight $\sigma_z^2/\sigma_{r_n}^2$
- Principled quantification of task ambiguity
- Only marginal computational overhead
- Compatible with existing NP architectures



Experiments

6

	PB/det.		VI		MC	
	BA	MA (CNP)	BA	MA (LP-NP)	BA	MA
RBF GP	1.37 ± 0.15	0.94 ± 0.04	1.40 ± 0.04	0.45 ± 0.12	1.62 ± 0.05	1.07 ± 0.05
Weakly Periodic GP	1.13 ± 0.08	0.76 ± 0.02	0.89 ± 0.03	0.07 ± 0.14	1.30 ± 0.06	0.85 ± 0.04
Matern-5/2 GP	-0.50 ± 0.07	-0.68 ± 0.01	-0.79 ± 0.01	-1.09 ± 0.10	-0.33 ± 0.01	-0.90 ± 0.15
Furuta Dynamics	7.50 ± 0.27	7.06 ± 0.12	7.32 ± 0.18	5.57 ± 0.21	8.25 ± 0.33	7.55 ± 0.24



ICLR 2021 | M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, Bayesian Context Aggregation for Neural Processes © Robert Bosh GmbH 2021. Al right reserved, also regarding any disposal, exploitation, reproduction, as well as in the event of applications for industrial property rights.

