What Matters For Meta-Learning Vision Regression Tasks?

Ning Gao[1,2] Hanna Ziesche[1] Ngo Anh Vien[1] Michael Volpp[2] Gerhard Neumann[2]

[1]Bosch Center for Artificial Intelligence [2]Autonomous Learning Robots, KIT

Code & Model

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

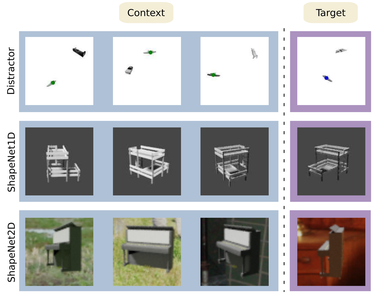
Problem

Goal: Exploit meta-learning algorithms on image-based regression tasks.

Contributions:

- Create three cross-category level vision regression tasks in the meta-learning domain, where predictions are conducted on novel objects from unseen categories based on few-shot context information.
- Exhaustive evaluation of different meta-learning approaches.
- Analysis of different techniques w.r.t. meta-learning overfitting.
- A simple and effective functional contrastive learning (FCL) over the task representations in Conditional Neural Processes (CNPs)

Task Design



- Predict the position of the queried object, where the queried object is identified by giving the position in the context images (Distractor).
- Predict the azimuth angle requires to identify the canonical pose of each specific object from the context set (ShapeNet1D).
- Predict both azimuth and elevation angle of the object with random background (ShapeNet2D).

Problem Formulation

Assume all tasks are under the same distribution $p(\mathcal{T})$, each task $\mathcal{T}_i$ includes a context set $\mathcal{D}_C^i = \{(x_{C,1}, y_{C,1}), ..., (x_{C,K}, y_{C,K})\}_i$ and a target set $\mathcal{D}_T^i = \{(x_{T,1}, y_{T,1}), ..., (x_{T,M}, y_{T,M})\}_i$ where K and M are the number of samples in each set. Training dataset is denoted as $\mathcal{D} = \{\mathcal{D}_C^i, \mathcal{D}_T^i\}_{i=1}^N$ where N is the number of tasks sampled for training. During inference, the model is tested on a new task $\mathcal{T}^* \sim p(\mathcal{T})$ given a small context set, from which it has to infer a new function $f^* : (\mathcal{D}_C^*, x_T^*) \rightarrow \hat{y}_T^*$

Meta-Overfitting

Memorization Overfitting

$$z^t = h_\theta(\mathcal{D}_C^t);\ \hat{y}_t = g_\phi(x^t, z^t) \rightarrow \hat{y}_t = g_\phi(x^t), t \in \mathcal{T}_{train}$$

Learner Overfitting

$$\hat{y}_t = g_\phi(x^t, z^t), t \in \mathcal{T}_{train};\ \hat{y}_t^* \neq g_\phi(x^{t^*}, z^{t^*}), t^* \in \mathcal{T}_{test}$$

Techniques avoind overfitting

Data Augmentation (DA): Affine, Dropout, Crop, Contrast, Brightness, Blur.
Domain Randomization (DR): regenerate background images during training for ShapeNet2D.
Task Augmentation (TA): add randomness to each task, encourage the meta-learner to learn non-trivial solutions instead of memorization.

$$D_C^{(t)} = \{x_{C,i}^{(t)},\ y_{C,i}^{(t)} + \epsilon^{(t)}\}_{i=1}^K$$

$$D_T^{(t)} = \{x_{T,i}^{(t)},\ y_{T,i}^{(t)} + \epsilon^{(t)}\}_{i=1}^M$$

where $\epsilon^{(t)}$ is sampled from a discrete set for each task
Meta Regularization (MR): regularization on the meta-parameters θ of the neural networks.

$$\mathcal{L} = \mathcal{L}_O + \beta D_{\mathrm{KL}}(q(\theta; \theta_\mu, \theta_\sigma)||r(\theta))$$

where $\mathcal{L}_O$ denotes the original loss function defined individually in Distractor and pose estimation.

Functional Contrastive Learning

CNPs only consider permutation-invariant task representations over the context order:

$$h_\theta\{\square\square\square\square\} = h_\theta\{\square\square\square\square\}$$

We consider one step further on the connections among different sets and tasks, namely yielding closed representations from sets of the same task and compelling the representations from different tasks over the functional space. This also helps to learn consistent representation during training:

$$h_\theta\{\underbrace{\square\square\square\square}_{t_1}\} \cong h_\theta\{\underbrace{\square\square\square\square}_{t_1}\} \qquad h_\theta\{\underbrace{\square\square\square\square}_{t_1}\} \neq h_\theta\{\underbrace{\square\square\square\square}_{t_2}\}$$

To achieve this, we employ the contrastive learning loss over the functional space between context and target sets:

$$\mathcal{L}_{\mathrm{FCL}} = -\frac{2}{N}\sum_{t=1}^{N}\log\frac{\exp(\mathrm{sim}(z_C^{(t)} \cdot z_T^{(t)})/\tau)}{D(z_C^{(t)})D(z_T^{(t)})}$$

where N denotes the number of tasks per batch, $D(z_i^t)$ sums the similarity of all positive and negative pairs for z_i^t:

$$D(z_i^t) = \sum_{k=1}^{N}\sum_{j\in\{C,T\}} 1_{[\{k\neq t\}\vee\{j\neq i\}]}\exp(\frac{\mathrm{sim}(z_i^t \cdot z_j^k)}{\tau})$$

where $1_{[\{k\neq t\}\vee\{j\neq i\}]} \in \{0, 1\}$ is an indicator evaluating to 1 only if the representations are sampled from different tasks or different sets.

Experiments & Results

Datasets:

- Distractor includes 12 object categories from ShapeNetCoreV2, where each category includes 1000 randomly sampled objects.
- Pascal1D contains 65 objects from 10 categories. 50 objects are randomly selected for training and the other 15 objects for testing.
- ShapeNet1D and ShapeNet2D include 30 categories. 27 of them are used during training and intra-category (IC) evaluation, the other 3 categories are used for cross-category (CC) evaluation.

Quantitative results on new toy tasks and prior Pascal1D:

Methods	MAML	CNP (Mean)	CNP (CA)
No Aug	1.69 (0.22)	5.28 (0.51)	4.66 (0.74)
MR	1.90 (0.27)	2.96 (0.21)	3.33 (0.27)
TA	**1.02 (0.06)**	**1.98 (0.22)**	**1.36 (0.25)**
DA	2.10 (0.09)	3.69 (0.13)	2.90 (0.03)
TA+DA	1.31 (0.14)	2.29 (0.19)	1.77 (0.33)

Prediction error on Pascal1D.

Methods	MAML	CNP (Max)	CNP (CA)
No Aug	25.27	14.97 (0.37)	8.19 (0.30)
	21.63	18.09 (0.21)	9.13 (0.18)
MR	13.23	12.71 (0.26)	8.87 (0.36)
	16.55	14.77 (0.35)	8.43 (0.39)
TA	23.01	10.89 (0.27)	7.92 (0.25)
	20.59	14.43 (0.55)	9.18 (0.50)
DA	14.69	8.64 (0.21)	6.24 (0.15)
	16.02	9.87 (0.35)	6.54 (0.19)
TA+DA	17.96	7.66 (0.18)	**5.81 (0.23)**
	18.79	8.66 (0.19)	**6.23 (0.12)**
TA+DA+FCL	–	7.82 (0.08)	6.44 (0.36)
	–	8.84 (0.04)	6.74 (0.20)
TA+DA+MR	13.45	10.54 (0.37)	8.28 (0.17)
	14.44	10.76 (0.30)	8.04 (0.10)

Prediction error(°) on ShapeNet1D.

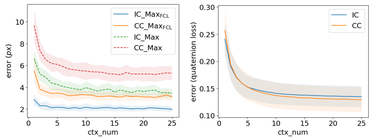
Methods	Mean	Max	BA	CA	$\mathrm{Max}_{\mathrm{FCL}}$
No Aug	6.02	5.11	4.63	5.13	**3.70**
	6.89	6.17	5.91	6.39	**4.61**
DA	2.67	2.45	2.44	2.65	**2.00**
	4.10	3.75	3.97	4.08	**3.05**
TA	6.29	6.18	6.33	6.32	**5.45**
	7.19	7.04	7.02	7.02	**6.66**
TA+DA	3.20	3.09	2.65	3.05	**2.60**
	6.07	5.14	4.67	4.98	**3.90**

Prediction error (pixel) of euclidean distance in the 2D image plane for Distractor.

Methods	IC ($1e^{-2}$)	CC ($1e^{-2}$)
None	38.33 (0.33)	39.81 (0.31)
DR	18.67 (0.13)	20.05 (0.12)
DR+MR	27.89 (0.61)	28.99 (0.46)
$\mathrm{DR+TA_{azi}}$	16.94 (0.13)	18.42 (0.26)
$\mathrm{DR+TA_{azi+ele}}$	16.62 (0.12)	17.76 (0.35)
DA	19.32 (0.09)	17.98 (0.09)
DR+DA	14.26 (0.09)	13.91 (0.14)
$\mathrm{DR+DA+TA_{azi+ele}}$	14.12 (0.14)	13.59 (0.10)
$\mathrm{DR+DA+TA_{azi+ele}}$ + FCL	**14.01 (0.09)**	**13.32 (0.18)**

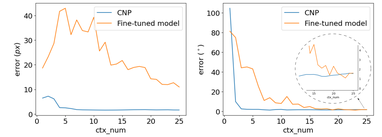
Prediction error on ShapeNet2D.

Accuracy vs context number:



(a) Prediction error (pixel) vs context number on Distractor and (b) ShapeNet2D.

CNPs vs finetuned classical model:



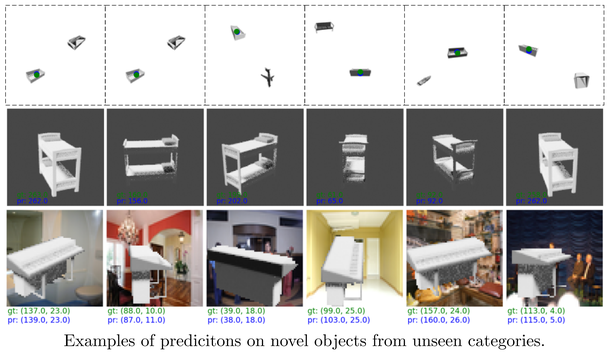
(a) Comparison between CNP (Max) and finetuned classical model on Distractor and (b) ShapeNet1D.

Analysis of data efficiency:

Methods	$\mathrm{CA_S}$	$\mathrm{CA_M}$	$\mathrm{CA_L}$	$\mathrm{Max_S}$	$\mathrm{Max_M}$	$\mathrm{Max_L}$
No Aug	18.60 (0.78)	12.08 (0.44)	8.19 (0.30)	30.44 (0.82)	18.86 (0.34)	14.97 (0.37)
	19.95 (1.08)	12.62 (0.87)	9.13 (0.18)	30.59 (1.14)	21.78 (0.47)	18.09 (0.21)
TA	18.69 (0.87)	10.70 (0.98)	7.92 (0.25)	21.67 (0.66)	13.69 (0.27)	10.89 (0.27)
	19.24 (0.79)	12.05 (0.73)	9.18 (0.50)	23.60 (0.88)	16.76 (0.62)	14.43 (0.55)
TA+DA	7.86 (0.21)	6.32 (0.11)	**5.81 (0.23)**	11.00 (0.16)	8.23 (0.34)	7.66 (0.18)
	7.49 (0.35)	6.48 (0.41)	**6.23 (0.12)**	12.98 (0.48)	9.65 (0.40)	8.66 (0.19)

Performance on ShapeNet1D using small (S), medium (M) and large (L) training dataset sizes for CNP.

Visualizations:



Examples of predicitons on novel objects from unseen categories.

Key takeaways:

- Prior work on Pascal1D uses inappropriate loss function for training, also Pascal1D lacks diverse object variations.
- CNPs outperform MAML with notable data/training efficiency with increasing task diversity.
- DA alleviates both types of overfitting while TA alleviates memorization overfitting but requires tailored design.
- CNPs surpasses fine-tuned models especially on the few-shot domain.
- Use max aggregation for non-positional encoding tasks and cross attention (CA) for object-centric tasks with positional information, mean aggregation shows poor performance.
- FCL can alleviate overfitting and increase the performance but requires finetuning the temperature term. In our toy tasks, using a small temperature value and FCL between context and target set normally gets good performance.