X-Model: Improving Data Efficiency in Deep Learning with a Minimax Model

Ximei Wang, Xinyang Chen, Jianmin Wang and Mingsheng Long (✉)

School of Software, BNRist, Tsinghua University

wxm17@mails.tsinghua.edu.cn

International Conference on Learning Representations (ICLR), 2022

Goal: Improving Data Efficiency in Deep Learning

- **Methods in classificattion setup adopt human intuitions**:
  - low density separation
  - cluster assumptions
  - pseudo labeling strategies
  - ...
- **Methods in regression setup focus on shallow learning**:
  - k-nearest neighbor (kNN)
  - decision tree
  - Gaussian Process
  - ...
- How to improve data efficiency for both classification and regression setups?

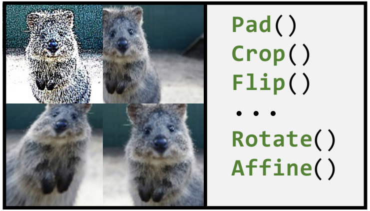
Can we further enhance model stochasticity as data stochasticity?

- **Type 1: Encourage invariance to data stochasticity**
  - consistency regularization to local input perturbations
  - Π-model, FixMatch, Unsupervised Data Augmentation, ...

- **Type 2: Encourage invariance to model stochasticity**
  - difference penalty for predictions of models generated from different dropout, initialization, exponentially averaged history models
  - Π-model, COREG, Mean Teacher, ...

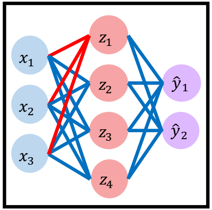
Can we further enhance model stochasticity as data stochasticity?

(a) **Data**

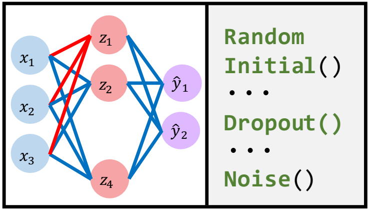


(b) **Data Stochasticity: Weak**

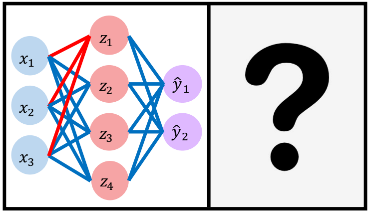
(c) **Data Stochasticity: Strong**



(d) **Model**



(e) **Model Stochasticity: Weak**



(f) **Model Stochasticity: Strong**

Table: Comparison among various methods for improving data efficiency in deep learning.

Method	stochasticity		setup	
	data	model	classification	regression
Pseudo Label	weak	✗	✓	✗
Entropy	✗	✗	✓	✗
VAT	weak	✗	✓	✓
Π-model	weak	weak	✓	✓
Data Distillation	weak	✗	✓	✗
Mean Teacher	weak	weak	✓	✓
UDA	strong	✗	✓	✗
FixMatch	strong	✗	✓	✗
Self-Tuning	strong	✗	✓	✗
X-model	strong	strong	✓	✓

Preliminary: Invariant to Data Stochasticity

- **Encourage invariance to data stochasticity**:

$$\min_{\theta,\phi} L_{\mathrm{data}}(\boldsymbol{x},\mathcal{U}) = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{U}}\ \ell\left((\phi\circ\theta)(\mathrm{aug}_1(\boldsymbol{x}_i)),\ (\phi\circ\theta)(\mathrm{aug}_2(\boldsymbol{x}_i))\right), \tag{1}$$

- Denote a labeled dataset $\mathcal{L} = \left\{\left(\boldsymbol{x}_i^L, \boldsymbol{y}_i^L\right)\right\}_{i=1}^{n_L}$ with n_L samples $\left(\boldsymbol{x}_i^L, \boldsymbol{y}_i^L\right)$
- Denote an unlabeled dataset $\mathcal{U} = \left\{\left(\boldsymbol{x}_i^U\right)\right\}_{i=1}^{n_U}$ with n_U unlabeled samples.
- The size n_L of $\mathcal{L}$ is much smaller than that n_U of $\mathcal{U}$ and the label ratio is $n_L/(n_L+n_U)$.
- Denote θ the feature generator network, and ϕ the successive task-specific head network.

Preliminary: Invariant to Model Stochasticity

- **Encourage invariance to model stochasticity**

$$\min_{\theta,\phi} L_{\text{model}}(\boldsymbol{x},\mathcal{U}) = \mathbb{E}_{\boldsymbol{x}_i\in\mathcal{U}}\ \ell\left((\phi_t\circ\theta_t)(\boldsymbol{x}_i),\ (\phi'_{t-1}\circ\theta'_{t-1})(\boldsymbol{x}_i)\right), \tag{2}$$

- $(\phi_t\circ\theta_t)$, $(\phi'_{t-1}\circ\theta'_{t-1})$ are the current model and the exponential moving averaged model
- $\phi'_t = \alpha\phi'_{t-1} + (1-\alpha)\phi_t, \theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$ where α is a smoothing coefficient hyperparameter.

Data Stochasticity meets Model Stochasticity

- The same feature extractor θ and two different task-specific heads ϕ_1 and ϕ_2:

$$\begin{aligned}\widehat{\boldsymbol{y}}_{i,1} &= (\phi_1 \circ \theta)(\mathrm{aug}_1(\boldsymbol{x}_i)) \\ \widehat{\boldsymbol{y}}_{i,2} &= (\phi_2 \circ \theta)(\mathrm{aug}_2(\boldsymbol{x}_i)),\end{aligned} \tag{3}$$

- For each example $\boldsymbol{x}_i$ in the labeled dataset $\mathcal{L} = \left\{\left(\boldsymbol{x}_i^L, \boldsymbol{y}_i^L\right)\right\}_{i=1}^{n_L}$,

$$L_s(\boldsymbol{x}, \mathcal{L}) = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{L}}\ \ell_s\left(\widehat{\boldsymbol{y}}_{i,1},\ \boldsymbol{y}_i\right) + \ell_s\left(\widehat{\boldsymbol{y}}_{i,2},\ \boldsymbol{y}_i\right), \tag{4}$$

- For each example in the unlabeled dataset $\mathcal{U} = \left\{\left(\boldsymbol{x}_i^U\right)\right\}_{i=1}^{n_U}$,

$$L_u(\boldsymbol{x}, \mathcal{U}) = \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{U}}\ \ell_u\left[(\phi_1 \circ \theta)(\mathrm{aug}_1(\boldsymbol{x}_i)),\ (\phi_2 \circ \theta)(\mathrm{aug}_2(\boldsymbol{x}_i))\right], \tag{5}$$

Enhance Model Stochasticity via a Minimax Model

- Only minimizing causes a degeneration problem and provides little meaningful information

$$L_{\mathrm{s}}(\boldsymbol{x}, \mathcal{L}) + \eta L_{\mathrm{u}}(\boldsymbol{x}, \mathcal{U})$$

- Enhance model stochasticity via a minimax model

$$\begin{aligned} \widehat{\theta} &= \arg\min_{\theta} \quad L_s(\boldsymbol{x}, \mathcal{L}) + \eta L_u(\boldsymbol{x}, \mathcal{U}), \\ (\widehat{\phi}_1, \widehat{\phi}_2) &= \arg\min_{\phi_1, \phi_2} \quad L_s(\boldsymbol{x}, \mathcal{L}) - \eta L_u(\boldsymbol{x}, \mathcal{U}), \end{aligned} \tag{6}$$

Results on Regression Setup

Table 2: MAE (↓) on tasks of Position X, Position Y and Scale in *dSprites-Scream* (ResNet-18).

Label Ratio	1%				5%				20%				50%			
Method	Scale	X	Y	**All**	Scale	X	Y	**All**	Scale	X	Y	**All**	Scale	X	Y	**All**
Only Labeled Data	.130	.073	.075	.277	.072	.036	.035	.144	.051	.030	.028	.108	.046	.026	.025	.097
VAT (Miyato et al., 2016)	.067	.042	.038	.147	.046	.028	.034	.109	.045	.024	.029	.098	.037	.027	.020	.084
Π-model (Laine & Aila, 2017)	.084	.035	.035	.154	.058	.031	.025	.114	.045	.024	.023	.092	.040	.021	.021	.082
Data Distillation (Radosavovic et al., 2017)	.066	.039	.033	.138	.045	.027	.031	.104	.043	.023	.026	.092	.037	.023	.021	.081
Mean Teacher (Tarvainen & Valpola, 2017)	.062	.035	.037	.134	.045	.024	.033	.103	.042	.023	.024	.089	.038	.021	.020	.079
UDA (Xie et al., 2020)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
FixMatch (Sohn et al., 2020)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Self-Tuning (Wang et al., 2021)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
χ**-model** (w/o minimax)	.080	**.021**	.024	.125	.044	.029	.028	.101	.040	.017	.021	.077	**.030**	.027	.018	.074
χ**-model** (w/o data aug.)	.074	.025	**.023**	.119	.045	.026	**.022**	.093	.037	.019	**.017**	.073	.038	.018	**.017**	.074
χ**-model** (ours)	**.061**	.030	.024	**.115**	**.044**	**.023**	.025	**.092**	**.037**	**.014**	.021	**.072**	.032	**.018**	.018	**.068**

Results on Classification Setup

Table 6: Error rates (%) $\downarrow$ of classification on *CIFAR-100* (WRN-28-8).

Method	400 labels	2500 labels	10000 labels
Pseudo-Labeling (Lee, 2013)	-	57.38$\pm$0.46	36.21$\pm$0.19
MC Dropout (Gal & Ghahramani, 2016)	-	58.27$\pm$0.54	38.36$\pm$0.19
Deep Co-Training (Qiao et al., 2018)	-	53.38$\pm$0.61	34.63$\pm$0.14
Π-Model (Laine & Aila, 2017)	-	57.25$\pm$0.48	37.88$\pm$0.11
MME (Saito et al., 2019)	-	47.40$\pm$1.75	32.54$\pm$0.81
Mean Teacher (Tarvainen & Valpola, 2017)	-	53.91$\pm$0.57	35.83$\pm$0.24
MixMatch (Berthelot et al., 2019)	67.61$\pm$1.32	39.94$\pm$0.37	28.31$\pm$0.33
UDA (Xie et al., 2020)	59.28$\pm$0.88	33.13$\pm$0.22	24.50$\pm$0.25
ReMixMatch (Berthelot et al., 2020)	**44.28**$\pm$2.06	27.43$\pm$0.31	23.03$\pm$0.56
FixMatch (Sohn et al., 2020)	48.85$\pm$1.75	28.29$\pm$0.11	22.60$\pm$0.12
Meta Pseudo Labels (Pham et al., 2021)	48.18$\pm$1.29	27.31$\pm$0.24	22.02$\pm$0.18
Self-Tuning (Wang et al., 2021)	54.74$\pm$0.35	42.08$\pm$0.43	21.75$\pm$0.27
χ-model	47.21$\pm$1.54	**27.11**$\pm$0.65	**20.98**$\pm$0.33

Summary

- We propose the X-model that jointly encourages the invariance to *data and model stochasticity* to improve data efficiency for both classification and regression setups.
- We make the X-model play a minimax game between the feature extractor and task-specific heads to further enhance invariance to model stochasticity.
- Extensive experiments verify the superiority of the X-model among various tasks, from an age estimation task to a dense-value prediction task of keypoint localization, a 2D synthetic and a 3D realistic dataset, as well as a multi-category object recognition task.