POSTECH
INDUSTRIAL AND MANAGEMENT ENGINEERING, POSTECH

SWAD: Domain Generalization by Seeking Flat Minima

포항공과대학교 산업경영공학과

Stochastic Systems Lab

Jongwon Kim

May 16, 2022

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, Sungrae Park, NeurIPS 2021

STOCHASTIC SYSTEMS LAB.

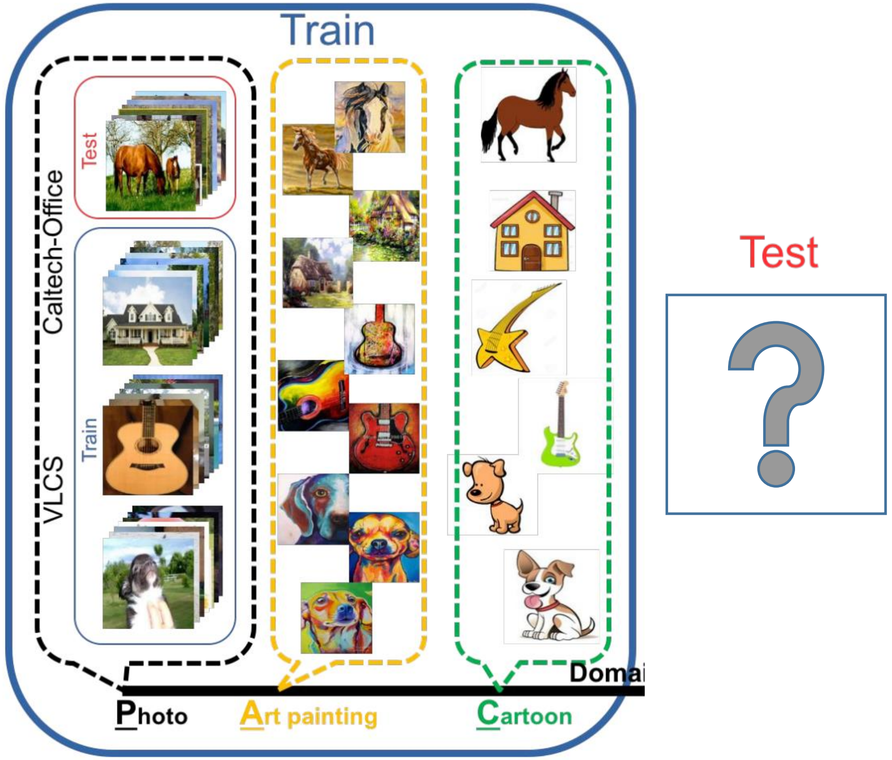
Contents

STOCHASTIC SYSTEMS LAB.

Problem statement

Why is domain generalization?

We have multi-domain training data.
We don't know the test task.

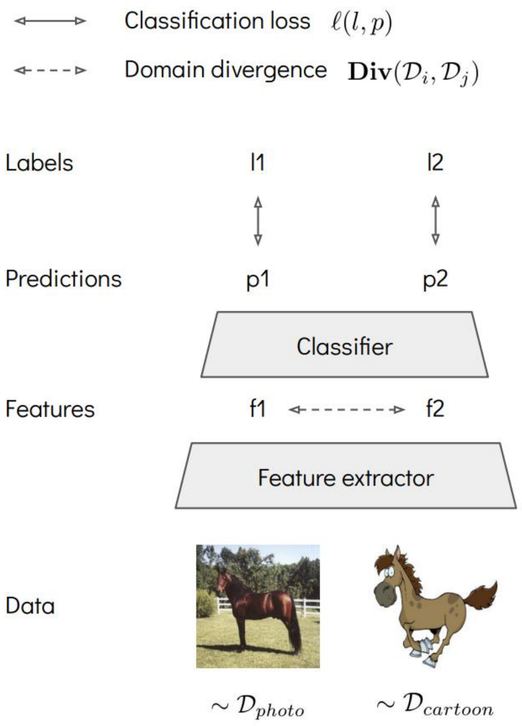


The figure is adopted from [1]

Previous studies

Main idea

Find **domain invariant** representation!



Domain divergence can be minimized implicitly, *e.g.*, by data augmentation.

Limitations of traditional domain generalization methods in the practical perspective

Degrade in-domain performance
Degrade training or inference speed
Be restricted to a task (e.g. classification) or domain (e.g. vision)
Require domain labels

SWAD: Domain Generalization by Seeking Flat Minima

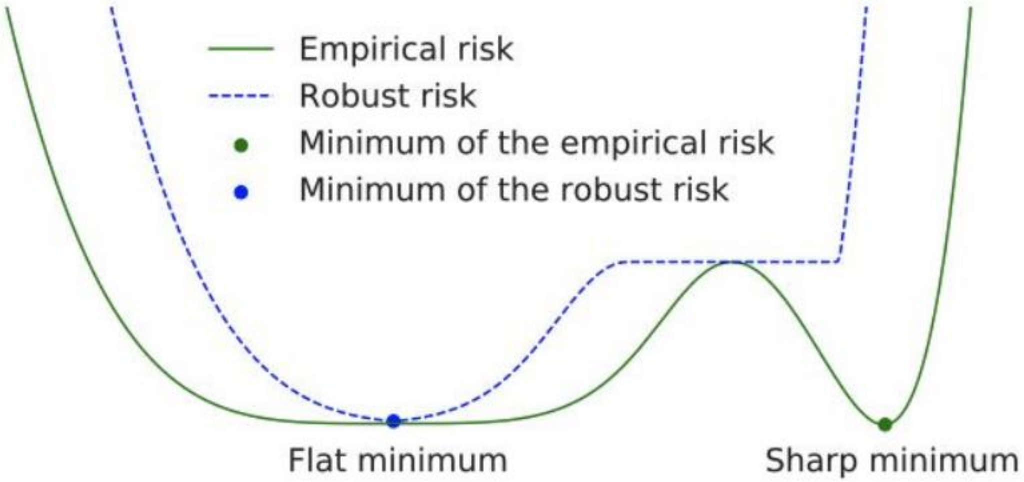
Main idea

Find flat minima!

θ : model parameter
ε : Loss function
D : Training data (Domain)
γ : robust parameter

$$\hat{\mathcal{E}}_{\mathcal{D}}^{\gamma}(\theta) := \max_{\|\Delta\| \leq \gamma} \hat{\mathcal{E}}_{\mathcal{D}}(\theta + \Delta)$$

$$\hat{\theta}^{\gamma} := \arg\min_{\theta} \hat{\mathcal{E}}_{\mathcal{D}}^{\gamma}(\theta)$$



SWAD: Domain Generalization by Seeking Flat Minima

Main idea

Find **flat minima**!

$$\hat{\theta}^{\gamma} := \arg\min_{\theta} \hat{\mathcal{E}}_{\mathcal{D}}^{\gamma}(\theta)$$

Objective function

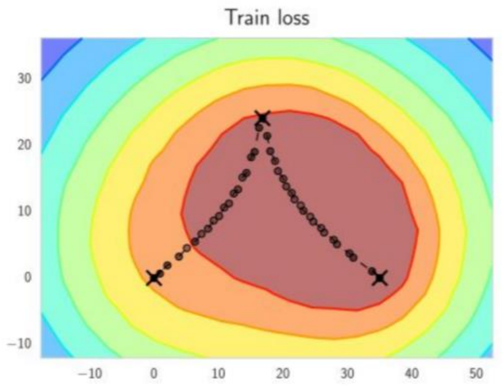
the optimum of robust risk

domain divergence

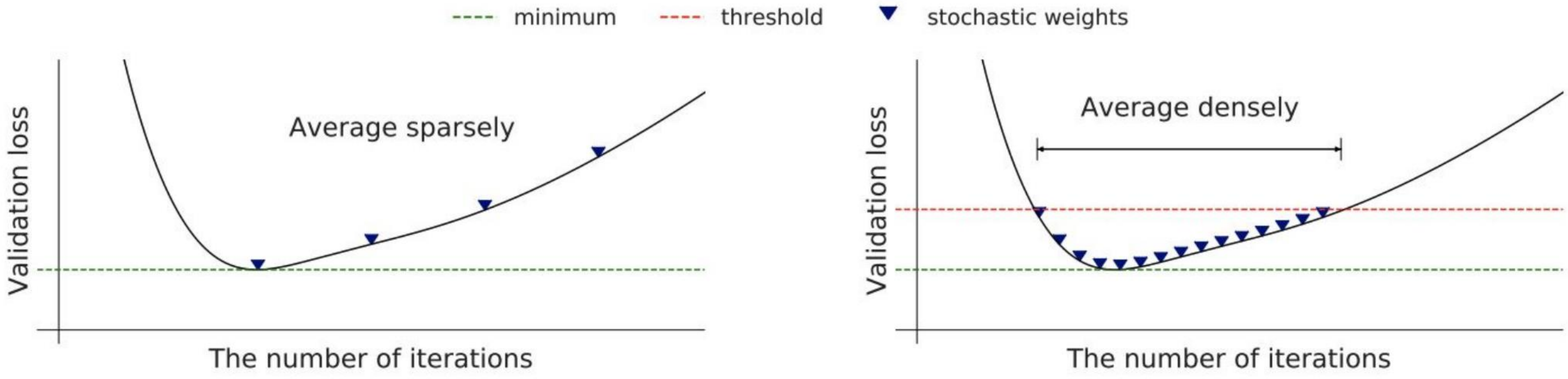
$$\underbrace{\mathcal{E}_{\mathcal{T}}(\hat{\theta}^{\gamma}) - \min_{\theta'} \mathcal{E}_{\mathcal{T}}(\theta')}_{\text{DG gap in test domain } T} \leq \underbrace{\hat{\mathcal{E}}_{\mathcal{D}}^{\gamma}(\hat{\theta}^{\gamma}) - \min_{\theta''} \hat{\mathcal{E}}_{\mathcal{D}}(\theta'')}_{\text{Gap between RRM and ERM in training domains } D} + \underbrace{\frac{1}{I}\sum_{i=1}^{I} \mathbf{Div}(\mathcal{D}_i, \mathcal{T})}_{\text{Domain divergence}} \underbrace{+ \max_{k\in[1,N]} \sqrt{\frac{v_k \ln(m/v_k) + \ln(2N/\delta)}{m}} + \sqrt{\frac{v \ln(m/v) + \ln(2/\delta)}{m}}}_{\text{Model capacity and sample size}}$$

SWAD: Domain Generalization by Seeking Flat Minima

Main idea



Adopted from [1]

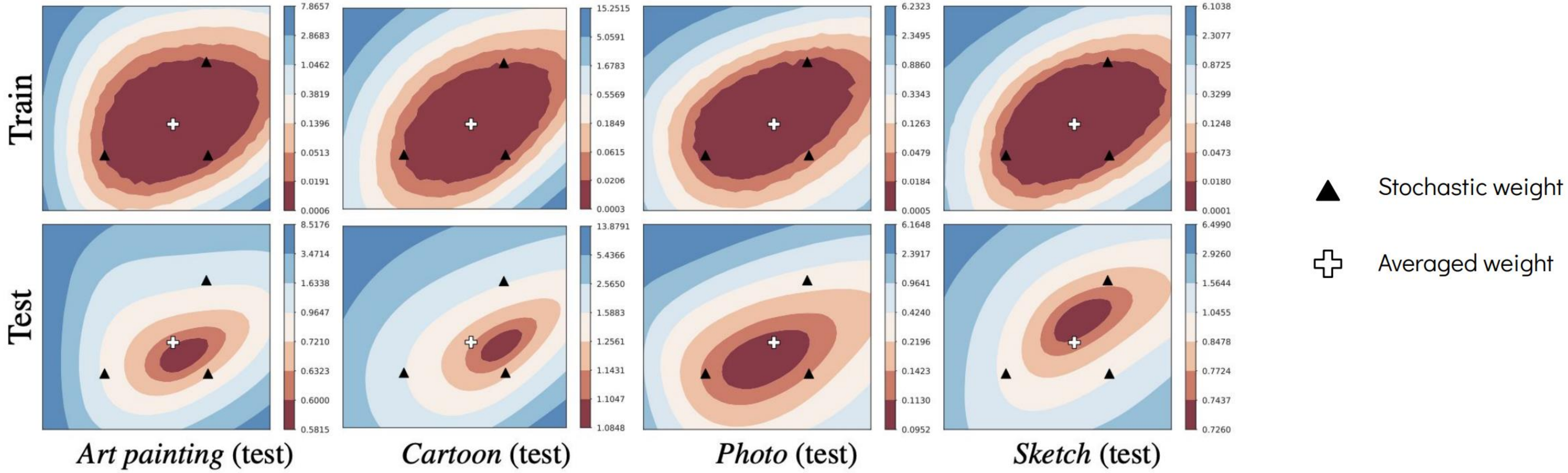


SWA

SWAD

SWAD: Domain Generalization by Seeking Flat Minima

Main idea



Results

Performance table

Comparison with conventional generalization methods

	Out-of-domain	In-domain
ERM	85.3±0.4	96.6±0.0
EMA	85.5±0.4(-)	97.0±0.1(↑)
SAM	85.5±0.1(-)	97.4±0.1(↑)
Mixup	84.8±0.3(-)	97.3±0.1(↑)
CutMix	83.8±0.4(↓)	97.6±0.1(↑)
VAT	85.4±0.6(-)	96.9±0.2(↑)
Π-model	83.5±0.5(↓)	96.8±0.2(↑)
SWA	85.9±0.1(↑)	97.1±0.1(↑)
SWAD	**87.1**±0.2(↑)	**97.7**±0.1(↑)

Comparison between conventional generalization methods on PACS

Comparison of flatness-aware solvers

Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
ERM (baseline)	85.5 ±0.2	77.5 ±0.4	66.5 ±0.3	46.1 ±1.8	40.9 ±0.1	63.3
SAM	85.8 ±0.2	79.4 ±0.1	69.6 ±0.1	43.3 ±0.7	44.3 ±0.0	64.5
SWA	87.1 ±0.1	76.5 ±0.2	68.5 ±0.2	49.6 ±1.0	45.6 ±0.0	65.5
SWAD	**88.1** ±0.1	**79.1** ±0.1	**70.6** ±0.2	**50.0** ±0.3	**46.5** ±0.1	**66.9**

Previous studies

Self-supervised learning

Combination with domain divergence minimization

	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg. (Δ)	Minimize: Robust risk	Minimize: Domain div.
ERM	85.5 $\pm$0.2	77.5 $\pm$0.4	66.5 $\pm$0.3	46.1 $\pm$1.8	40.9 $\pm$0.1	63.3		
ERM + SWAD	88.1 $\pm$0.1	79.1 $\pm$0.1	70.6 $\pm$0.2	50.0 $\pm$0.3	46.5 $\pm$0.1	66.9 (+3.6)	✔	
CORAL	86.2 $\pm$0.3	78.8 $\pm$0.6	68.7 $\pm$0.3	47.6 $\pm$1.0	41.5 $\pm$0.1	64.5		✔
CORAL + SWAD	88.3 $\pm$0.1	78.9 $\pm$0.1	71.3 $\pm$0.1	51.0 $\pm$0.1	46.8 $\pm$0.0	67.3 (+2.8)	✔	✔

Thank you