Generalizing to Unseen Domains via Adversarial Data Augmentation

Authors: Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, Silvio Savarese
(IIT & Stanford)

Presenters: Liliang Ren and Zexing Xu
Outline

- Introduction
- Method
- Theoretical Analysis
- Experiment
- Conclusions & Future work
Motivation

- **Goal:** learn a system that can perform uniformly well across multiple populations

- **Method:** domain adaptation -> perform well on known target distributions

- **Challenge:** only a limited number of population sources
  unrealistic to assume known target distribution
Motivation

- Example: semantic segmentation problem for robot
- difficult to collect samples from diverse scenarios
Introduction

- Goal: find a method that can better generalize to unknown domains

- Problem Setting:
  - Training data only comes from a single source domain
  - Test on data from other domains
Contribution

- a worst-case formulation over data distributions that are near the source domain in the feature space
- an iterative procedure that augments the dataset with examples from a fictitious target domain
- The method learns models improve performance across a range of a priori unknown target domains.
Related Works

✓ Adversarial training ~ resistant to larger perturbations, minimax game

✗ Domain adaptation ~ No access to samples from the target distribution during training

✓ Domain generalization ~ unsupervised domain generalization

✓ Domain randomization ~ actually learning new data points
Our approach

Considering the worst-case problem around the source distribution, $P_0$:

$$\min_{\theta \in \Theta} \sup_{P: D(P, P_0) \leq \rho} \mathbb{E}_P[\ell(\theta; (X, Y))]$$

where $D$ is a distance metric of distributions, and rho is the hyperparameter. Applying the Lagrangian relaxation,

$$\min_{\theta \in \Theta} \sup_{P} \{\mathbb{E}_P[\ell(\theta; (X, Y))] - \gamma D_\theta(P, P_0)\}$$
Our approach

Defining the robust surrogate loss,

\[
\phi_\gamma(\theta; (x_0, y_0)) := \sup_{x \in X} \left\{ \ell(\theta; (x, y_0)) - \gamma c_\theta((x, y_0), (x_0, y_0)) \right\}
\]

where,

\[
c((z, y), (z', y')) := \frac{1}{2} \|z - z'\|_2^2 + \infty \cdot 1 \{y \neq y'\}
\]

is the mass moving cost in the semantic space.
Our approach

With the following lemma,

Lemma 1. Let $\ell : \Theta \times (X \times Y) \rightarrow \mathbb{R}$ be continuous. For any distribution $Q$ and any $\gamma \geq 0$,
$$\sup_P \{\mathbb{E}_P[\ell(\theta; (X, Y))] - \gamma D_\theta(P, Q)\} = \mathbb{E}_Q[\phi_\gamma(\theta; (X, Y))].$$

We can augment the dataset by iteratively find the $x$ that maximize the surrogate loss which is minimized by the model parameters.
Iterative Procedure

Algorithm 1 Adversarial Data Augmentation

Input: original dataset \( \{X_i, Y_i\}_{i=1,...,n} \) and initialized weights \( \theta_0 \)
Output: learned weights \( \theta \)

1: Initialize: \( \theta \leftarrow \theta_0 \)
2: for \( k = 1, ..., K \) do  
   \( \triangleright \) Run the minimax procedure \( K \) times
3:   for \( t = 1, ..., T_{\text{min}} \) do
4:     Sample \( (X_t, Y_t) \) uniformly from dataset
5:     \( \theta \leftarrow \theta - \alpha \nabla_\theta \ell(\theta; (X_t, Y_t)) \)
6:   Sample \( \{X_i, Y_i\}_{i=1,...,n} \) uniformly from the dataset
7:   for \( i = 1, ..., n \) do
8:     \( X_i^k \leftarrow X_i \)
9:     for \( t = 1, ..., T_{\text{max}} \) do
10:    \( X_i^k \leftarrow X_i^k + \eta \nabla_x \{ \ell(\theta; (X_i^k, Y_i)) - \gamma c_\theta( (X_i^k, Y_i), (X_i, Y_i)) \} \)
11:   Append \( (X_i^k, Y_i^k) \) to dataset
12: for \( t = 1, ..., T \) do
13: Sample \( (X, Y) \) uniformly from dataset
14: \( \theta \leftarrow \theta - \alpha \nabla_\theta \ell(\theta; (X, Y)) \)
Ensembles for classification

The hyperparameter, $\gamma$, are chosen with s number of different values to train s different models. Given a test sample, the model with the greatest softmax score is selected,

$$u^*(x) := \arg\max_{1 \leq u \leq s} \max_{1 \leq j \leq k} \theta_{c,j}^u g(\theta_f^u; x).$$
Theoretical Analysis

- Adaptive Data Augmentation:
  the augmented data points can be interpreted as Tikhonov regularized Newton-steps

- Data-dependent Regularization:
  robust surrogate roughly corresponds to a novel data-dependent regularization scheme on the SoftMax loss
Preliminaries

- **Tikhonov regularization**: a method of regularization of ill-posed problems. Also known as ridge regression.
- the loss function is given by
  \[
  \mathcal{L}(x) = \|Ax - b\|_2^2 + \|\Gamma x\|_2^2
  \]
- so the gradient of it is
  \[
  \nabla_x \mathcal{L}(x) = \nabla_x (\|Ax - b\|_2^2) + \nabla_x (\|\Gamma x\|_2^2) = (2A^T Ax - 2A^T b) + (2\Gamma^T \Gamma x).
  \]
- Also, the hessian is given by
  \[
  \text{Hess } \mathcal{L}(x) = 2(A^T A + \Gamma^T \Gamma).
  \]
Theoretical Analysis

Adaptive Data Augmentation:

We fix \( \theta \in \Theta, x_0 \in X, y_0 \in Y \), and consider an \( \epsilon \)-maximizer

\[
x^*_\epsilon \in \epsilon \text{-arg max} \left\{ \ell(\theta; (x, y_0)) - \gamma c_\theta((x, y_0), (x_0, y_0)) \right\}.
\]

We let \( z_0 := g(\theta_f; x_0) \in \mathbb{R}^p \), and abuse notation by using \( \ell(\theta; (z_0, y_0)) := \ell(\theta; (x_0, y_0)) \). In what follows, we show that the feature mapping \( g(\theta_f; x^*_\epsilon) \) satisfies

\[
g(\theta_f; x^*_\epsilon) = g(\theta_f; x_0) + \frac{1}{\gamma} \left( I - \frac{1}{\gamma} \nabla_z \ell(\theta; (z_0, y_0)) \right)^{-1} \nabla_z \ell(\theta; (z_0, y_0)) + O \left( \sqrt{\frac{\epsilon}{\gamma}} + \frac{1}{\gamma^2} \right).
\]

\[
=: \tilde{g}_{\text{newton}}(\theta_f; x_0)
\]

- the adversarially perturbed sample \( x^*_\epsilon \) is drawn from a fictitious target distribution
- the transported point in the semantic space corresponds to a Tikhonov regularized Newton-step
Theoretical Analysis

Assume sufficient smoothness

**Assumption 1.** There exists $L_0, L_1 > 0$ such that, for all $z, z' \in \mathbb{R}^p$, we have $|\ell(\theta; (z, y_0)) - \ell(\theta; (z', y_0))| \leq L_0 \|z - z'\|_2$ and $\|\nabla_{z} \ell(\theta; (z, y_0)) - \nabla_{z} \ell(\theta; (z', y_0))\|_2 \leq L_1 \|z - z'\|_2$.

**Assumption 2.** There exists $L_2 > 0$ such that, for all $z, z' \in \mathbb{R}^p$, we have $\|\nabla_{zz} \ell(\theta; (z, y_0)) - \nabla_{zz} \ell(\theta; (z', y_0))\| \leq L_2 \|z - z'\|_2$. 
Theoretical Analysis

Adaptive Data Augmentation:

The bound for $g(\theta_f; x_c^*)$

**Theorem 1.** Let Assumptions 1, 2 hold. If $\text{Im}(g(\theta_f; \cdot)) = \mathbb{R}^p$ and $\gamma > L_1$, then

$$
\|g(\theta_f; x_c^*) - \hat{g}_{\text{newton}}(\theta_f; x_0)\|_2^2 \leq \frac{2\epsilon}{\gamma - L_1} + \frac{L_2}{3(\gamma - L_1)} \left\{ \left( \frac{5L_0}{\gamma} \right)^3 + \left( \frac{L_0}{\gamma - L_1} \right)^3 + \left( \frac{2\epsilon}{\gamma} \right)^{3/2} \right\}.
$$
Proof Sketch of Theorem 1

\[ \frac{\gamma - L_1}{2} \| z^* - \hat{g}_{\text{newton}}(\theta f; x_0) \|_2^2 \leq \epsilon + h_2(z^*_\epsilon) - h(z^*_\epsilon) \]

\[ + h(\hat{g}_{\text{newton}}(\theta f; x_0)) - h_2 (\hat{g}_{\text{newton}}(\theta f; x_0)) \]

\[ \leq \epsilon + \frac{L_2}{6} \left( \| z_0 - z^* \|_2^3 + \| z_0 - \hat{g}_{\text{newton}}(\theta f; x_0) \|_2^3 \right) \]

\[ h_2 \text{ and } h \text{ are close by Taylor expansion} \]

\[ \ell(\theta; (z, y_0)) - \frac{\gamma}{2} \| z - z_0 \|_2^2 =: h(z) \]

\[ \ell_2(\theta; (z, y_0)) - \frac{\gamma}{2} \| z - z_0 \|_2^2 := h_2(z) \]

Lemma 2 ([29, Lemma 1]). Let \( f : \mathbb{R}^p \to \mathbb{R} \) have a L-Lipschitz Hessian so that for all \( z, z' \in \mathbb{R}^p \),

\[ \| \nabla_{zz} f(z) - \nabla_{zz} f(z') \| \leq L \| z - z' \|_2. \]

Then, for all \( z, z' \in \mathbb{R}^p \),

\[ \left| f(z') - f(z) - \nabla f(z)^\top (z' - z) - \frac{1}{2} (z' - z)^\top \nabla_{zz} f(z) (z' - z) \right| \leq \frac{L}{6} \| z' - z \|_2^2. \]
Proof Sketch of Theorem 1

\[ \frac{\gamma - L_1}{2} \| z_e^\ast - \hat{g}_{\text{newton}}(\theta_f; x_0) \|_2^2 \leq \epsilon + \frac{L_2}{6} \left( \| z_0 - z_e^\ast \|_2^3 + \| z_0 - \hat{g}_{\text{newton}}(\theta_f; x_0) \|_2^3 \right) \]

\[ \hat{g}_{\text{newton}}(\theta_f; x_0) \] attains the maximum in the problem

\[ \hat{g}_{\text{newton}}(\theta_f; x_0) = z_0 + \frac{1}{\gamma} \left( I - \frac{1}{\gamma} \nabla z \ell(\theta; (z_0, y_0)) \right)^{-1} \nabla z \ell(\theta; (z_0, y_0)) \]

\[ = \arg \max_{z \in \mathbb{R}^p} \left\{ \ell_2(\theta; (z, y_0)) - \frac{\gamma}{2} \| z - z_0 \|_2^2 : = h_2(z) \right\} \]

\[ \| z_0 - \hat{g}_{\text{newton}}(\theta_f; x_0) \|_2^3 \leq \left( \frac{1}{\gamma} \right)^3 \left( \frac{\gamma}{\gamma - L_1} \right)^3 L_0^3. \]
Proof Sketch of Theorem 1

$$\frac{\gamma - L_1}{2} \| z_\epsilon^* - \hat{g}_{\text{newton}}(\theta_f; x_0) \|^2 \leq \epsilon + \frac{L_2}{6} \left( \| z_0 - z_\epsilon^* \|^3_2 + \| z_0 - \hat{g}_{\text{newton}}(\theta_f; x_0) \|^3_2 \right)$$


Lemma 3. Let Assumption 1 hold and $\text{Im}(g(\theta_f; \cdot)) = \mathbb{R}^p$. Then,

$$\left\| z_\epsilon^* - z_0 - \frac{1}{\gamma} \nabla_z \ell(\theta; (z_0, y_0)) \right\|_2 \leq \frac{4L_0}{\gamma} + \sqrt{\frac{2\epsilon}{\gamma}}.$$
Theoretical Analysis

Data-dependent Regularization:

- under suitable conditions on the loss,
- the robust surrogate is roughly equivalent to data-dependent regularization

\[
\phi_\gamma(\theta; (x, y)) = \ell(\theta; (x, y)) + \frac{1}{\gamma} \left\| \theta_{c,y} - \sum_{j=1}^{m} p_j(\theta, x) \theta_{c,j} \right\|_2^2 + O \left( \frac{1}{\gamma^2} \right)
\]

- minimize the distance between “average estimated linear classifier” to the linear classifier corresponding to the true label y.

**Theorem 2.** If \( \text{Im}(g(\bar{\theta}_f; \cdot)) = \mathbb{R}^p \) and \( \gamma > L(\theta) \), the softmax loss (2) satisfies

\[
\frac{1}{\gamma + L(\theta)} \left\| \theta_{c,y} - \sum_{j=1}^{m} p_j(\theta, x) \theta_{c,j} \right\|_2^2 \leq \phi_\gamma(\theta, (x, y)) - \ell(\theta, (x, y)) \leq \frac{1}{\gamma - L(\theta)} \left\| \theta_{c,y} - \sum_{j=1}^{m} p_j(\theta, x) \theta_{c,j} \right\|_2^2
\]

\[L(\theta) := 2 \max_{1 \leq j' \leq m} \| \theta_{c,j'} \|_2 \sum_{j=1}^{m} \| \theta_{c,j} \|_2\]
Proof Sketch of Theorem 2

Claim 5. If $z \mapsto \nabla_z \ell(\theta; (z, y))$ is $L$-Lipschitz with respect to the $\| \cdot \|_2$-norm, then

$$
\frac{1}{\gamma + L} \left\| \nabla_z \ell(\theta; (z, y)) \right\|_2^2 \leq \phi_\gamma(\theta; (z, y)) - \ell(\theta; (z, y)) \leq \frac{1}{\gamma - L} \left\| \nabla_z \ell(\theta; (z, y)) \right\|_2^2
$$

$$
\phi_\gamma(\theta; (x, y)) = \sup_{z' \in \mathbb{R}^p} \left\{ \ell(\theta; (z', y)) - \frac{\gamma}{2} \| z - z' \|_2^2 \right\}
$$

$z \mapsto \nabla_z \ell(\theta; (z, y))$ is $L$-Lipschitz.
Experimental Setting

● **Digital Classification:**
  ○ Train on MNIST,
  ○ Test on MNIST-M, SVHN, SYN, USPS

● **Semantic scene segmentation:**
  ○ SYTHIA dataset
  ○ Train on one location (Highway, New York-like City and Old European Town)
  ○ Test on another location without ensembles
Results - Digital Classification

Constraints:

Gamma: ERM
Results - Digital Classification

K (Single Model):

Ridge Loss:

K (Ensembles):
Results - Semantic Segmentation

Train on *Highway*

Test on *NY – like city*

Train on *New york – like City*

Test on *Highway*

Test on *Old European Town*

Test on *Old European Town*
Conclusions and Future Work

- Propose a novel data augmentation technique that generalize better to unseen domains
- Our iterative procedure provides broad generalization behavior
- For future work, we will extend the ensemble methods by defining novel decision rules
Thanks for your time