Generalizing to Unseen Domains via Adversarial Data Augmentation (NeurIPS 2018)

Adversarial Reading Group

Notions

- Wasserstein distance

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

$$D_\theta(P,Q) := \inf_{M \in \Pi(P,Q)} \mathbb{E}_M[c_\theta((X,Y),(X',Y'))].$$

- Worst case problem around training distribution.

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{P : D(P,P_0) \le \rho} \mathbb{E}_P[\ell(\theta;(X,Y))].$$

- Relaxation

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

Algorithm

Algorithm 1 Adversarial Data Augmentation

Input: original dataset $\{X_i, Y_i\}_{i=1,\ldots,n}$ and initialized weights θ_0
Output: learned weights θ

Initialize: $\theta \leftarrow \theta_0$
for $k = 1, \ldots, K$ **do** ▷ Run the minimax procedure K times
    for $t = 1, \ldots, T_{\min}$ **do**
        Sample (X_t, Y_t) uniformly from dataset
        $\theta \leftarrow \theta - \alpha \nabla_\theta \ell(\theta; (X_t, Y_t))$
    Sample $\{X_i, Y_i\}_{i=1,\ldots,n}$ uniformly from the dataset
    for $i = 1, \ldots, n$ **do**
        $X_i^k \leftarrow X_i$
        for $t = 1, \ldots, T_{\max}$ **do**
            $X_i^k \leftarrow X_i^k + \eta \nabla_x \left\{ \ell(\theta; (X_i^k, Y_i)) - \gamma c_\theta((X_i^k, Y_i), (X_i, Y_i)) \right\}$
        Append (X_i^k, Y_i^k) to dataset
for $t = 1, \ldots, T$ **do**
    Sample (X, Y) uniformly from dataset
    $\theta \leftarrow \theta - \alpha \nabla_\theta \ell(\theta; (X, Y))$

Results (1)

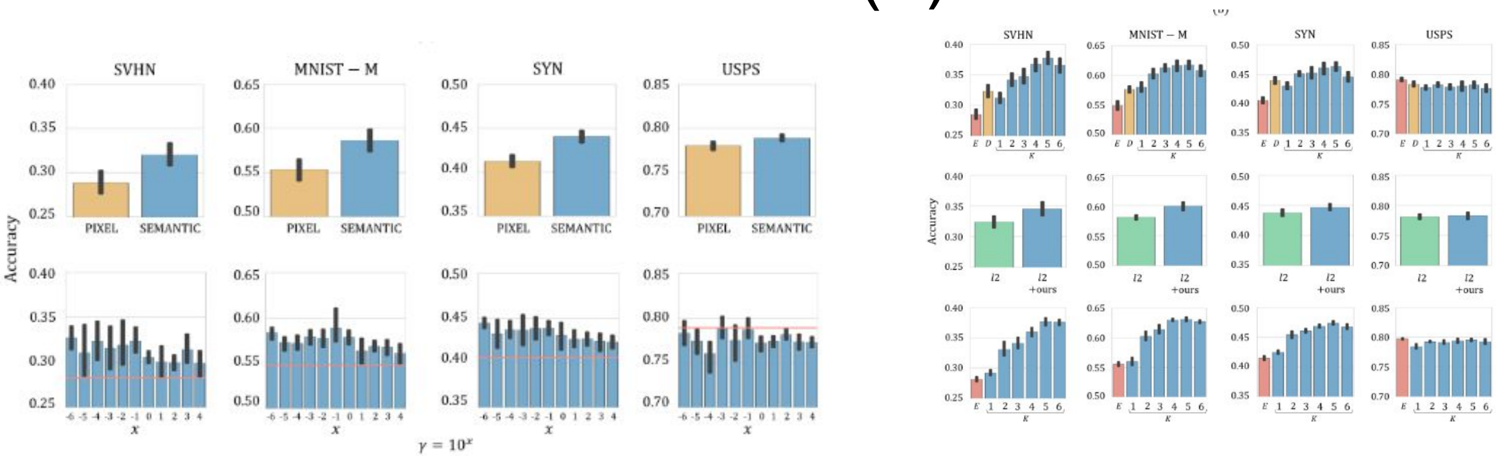

Figure 1. Results associated with models trained with 10,000 MNIST samples and tested on SVHN, MNIST-M, SYN and USPS (1^{st}, 2^{nd}, 3^{rd} and 4^{th} columns, respectively). *Panel (a), top:* comparison between distances in the pixel space (yellow) and in the semantic space (blue), with $\gamma = 10^4$ and $K = 1$. *Panel (a), bottom:* comparison between our method with $K = 2$ and different γ values (blue bars) and ERM (red line). *Panel (b), top:* comparison between our method with $\gamma = 1.0$ and different number of iterations K (blue), ERM (red) and Dropout [35] (yellow). *Panel (b), middle:* comparison between models regularized with ridge (green) and with ridge + our method with $\gamma = 1.0$ and $K = 1$ (blue). *Panel (b), bottom:* results related to the ensemble method, using models trained with our methods with different number of iterations K (blue) and using models trained via ERM (red). The reported results are obtained by averaging over 10 different runs; black bars indicate the range of accuracy spanned.

Results (2)

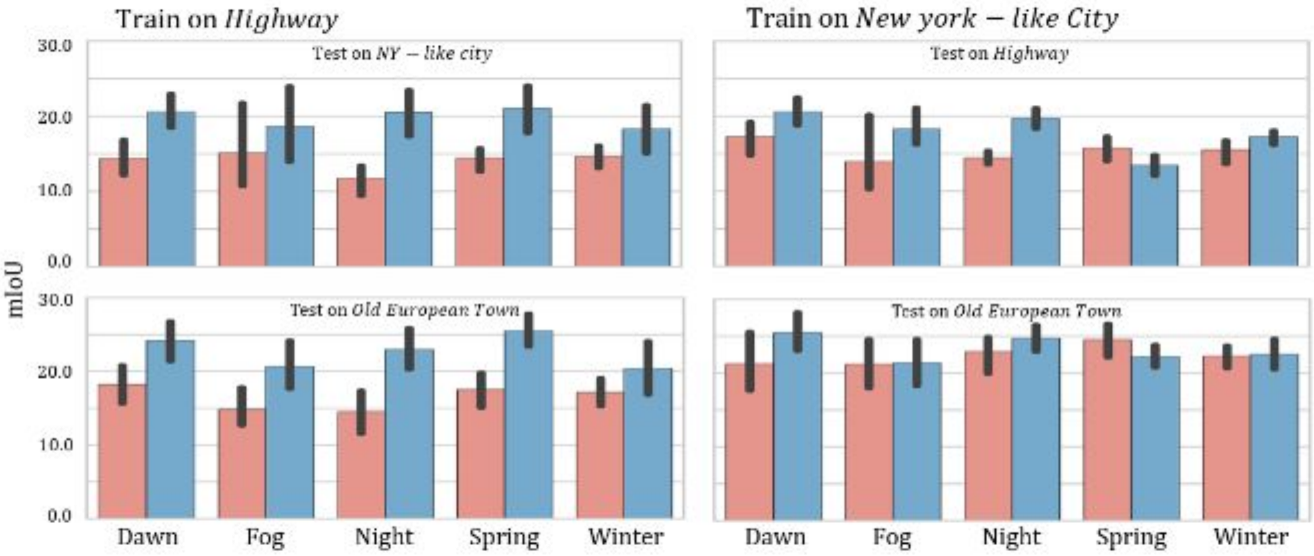

Figure 2. Results obtained with semantic segmentation models trained with ERM *(red)* and our method with $K = 1$ and $\gamma = 1.0$ *(blue)*. *Leftmost* panels are associated with models trained on *Highway*, *rightmost* panels are associated with models trained on *New York-like City*. Test datasets are *Highway*, *New York-like City* and *Old European Town*.

Questions?

Reading Group discussion

- Topics of interest
- Speaker ideas
- Project updates
- Getting involved