CSE 291: Domain Adaptation in Vision
Manmohan Chandraker

Lecture 13: H-Divergence Theory
Papers for Fri, Mar 13

• Domain Adaptive Faster R-CNN for Object Detection in the Wild
  • https://arxiv.org/abs/1803.03243

• Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection
  • https://arxiv.org/abs/1909.00597
Training DANN

• Optimize for three sets of parameters
  – Feature extractor $G_f$, label predictor $G_y$, domain discriminator $G_d$
H-divergence

- Consider the hypothesis class of functions $h: X \rightarrow \{0, 1\}$
- H-divergence between two domains

$$d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{x \sim D_S} [h(x) = 1] - \Pr_{x \sim D_T} [h(x) = 1] \right|$$

Empirical H-divergence converges to the true H-divergence

- $U, U'$: samples of size $m$ from distributions $D, D'$
- $d$: VC dimension of hypothesis space $\mathcal{H}$
- For $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$d_{\mathcal{H}}(D, D') \leq \hat{d}_{\mathcal{H}}(U, U') + 4 \sqrt{\frac{d \log(2m) + \log(\frac{2}{\delta})}{m}}$$

- For any sample $U, U'$, it is guaranteed true H-divergence not too high
Theory: computing $H$-divergence

- Can compute $H$-divergence for unlabeled domains
- Compute $H$-divergence by finding classifier that separates domains
  - Label each unlabeled source instance 0, unlabeled target instance 1
  - Train a classifier from hypothesis class $H$ to separate domains
  - $H$-divergence is related to the minimum classification error

  \[
  \hat{d}_H(U, U') = 2 \left( 1 - \min_{h \in H} \left[ \frac{1}{m} \sum_{x:h(x)=0} I[x \in U] + \frac{1}{m} \sum_{x:h(x)=1} I[x \in U'] \right] \right),
  \]

  - $I[x \in U]$ is binary indicator variable which is 1 when $x \in U$.
  - Roughly, $H$-divergence relates to the error of the best domain discriminator

- Minimizing error for most hypothesis classes is intractable
  - In practice, hypotheses trained to minimize convex upper bounds on error also well-approximate $H$-divergence
Theory: bounding risk in target domain

- Domain: a pair $(D, f)$ consisting of distribution $D$ on inputs $X$, with labeling function $f: X \rightarrow [0, 1]$
- We consider two domains: source $(D_S, f_S)$ and target $(D_T, f_T)$
- Hypothesis: a function $h: X \rightarrow \{0, 1\}$
- Error: probability that hypothesis $h$ disagrees with labeling $f$
  \[ \epsilon_S(h, f) = E_{x \sim D_S} [\|h(x) - f(x)\|] \]
- Source error (risk) of a hypothesis: \( \epsilon_S(h) = \epsilon_S(h, f_S) \)
- Empirical source error: \( \hat{\epsilon}_S(h) \)
- Similarly define target risk and empirical target error
Theory: symmetric difference hypothesis space

• Can compute H-divergence for unlabeled domains
• To use in bound, must compare error relative to other hypotheses
Theory: symmetric difference hypothesis space

- Can compute H-divergence for unlabeled domains
- To use in bound, must compare error relative to other hypotheses
- Symmetric difference hypothesis space $H \Delta H$
  - Set of hypotheses $g$ such that $g(x) = h(x) \oplus h'(x)$ for some $h, h' \in H$
  - Every hypothesis in $H \Delta H$ is set of disagreements between hypotheses in H
Theory: symmetric difference hypothesis space

- Can compute H-divergence for unlabeled domains
- To use in bound, must compare error relative to other hypotheses
- Symmetric difference hypothesis space $H\Delta H$
  - Set of hypotheses $g$ such that $g(x) = h(x) \oplus h'(x)$ for some $h, h' \in H$
  - XOR function
  - Every hypothesis in $H\Delta H$ is set of disagreements between hypotheses in H
- For any hypotheses $h, h' \in H$: $|\epsilon_S(h, h') - \epsilon_T(h, h')| \leq \frac{1}{2}d_{H\Delta H}(\mathcal{D}_S, \mathcal{D}_T)$
Theory: symmetric difference hypothesis space

• Can compute H-divergence for unlabeled domains
• To use in bound, must compare error relative to other hypotheses

• Symmetric difference hypothesis space $H \Delta H$
  – Set of hypotheses $g$ such that $g(x) = h(x) \oplus h'(x)$ for some $h, h' \in H$
  – Every hypothesis in $H \Delta H$ is set of disagreements between hypotheses in $H$

• For any hypotheses $h, h' \in H$: $|\epsilon_S(h, h') - \epsilon_T(h, h')| \leq \frac{1}{2} d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T)$

• Follows by definition of $H \Delta H$-divergence

$$d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h, h' \in \mathcal{H}} |\Pr_{x \sim \mathcal{D}_S}[h(x) \neq h'(x)] - \Pr_{x \sim \mathcal{D}_T}[h(x) \neq h'(x)]|$$

$$= 2 \sup_{h, h' \in \mathcal{H}} |\epsilon_S(h, h') - \epsilon_T(h, h')| \geq 2|\epsilon_S(h, h') - \epsilon_T(h, h')|.$$
Theory: bounding target risk

- Ideal joint hypothesis: minimizes sum of source and target risks
  \[ h^* = \arg\min_{h \in \mathcal{H}} \epsilon_S(h) + \epsilon_T(h) \]

- Combined error of ideal joint hypothesis
  \[ \lambda = \epsilon_S(h^*) + \epsilon_T(h^*) \]

- Characterizes notion of adaptability
  - Domain adaptation possible only if \( \lambda \) is small
  - Some classifier must do well on both source and target domains
Theory: bounding risk in target domain

• Symmetric difference hypothesis space $H \Delta H$
  – Set of hypotheses $g$ such that $g(x) = h(x) \oplus h'(x)$ for some $h, h' \in H$
  – Every hypothesis in $H \Delta H$ is set of disagreements between hypotheses in $H$

• Combined error of ideal joint hypothesis

$$\lambda = \epsilon_S(h^*) + \epsilon_T(h^*)$$
Theory: bounding risk in target domain

• Symmetric difference hypothesis space $H \Delta H$
  – Set of hypotheses $g$ such that $g(x) = h(x) \oplus h'(x)$ for some $h, h' \in H$
  – Every hypothesis in $H \Delta H$ is set of disagreements between hypotheses in $H$

• Combined error of ideal joint hypothesis

\[ \lambda = \varepsilon_S(h^*) + \varepsilon_T(h^*) \]

• Bound on target error obtainable in terms of $H \Delta H$-divergence

\[ \varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} \hat{d}_{H \Delta H}(U_S, U_T) + 4 \sqrt{\frac{2d \log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda \]

  – $U_S$, $U_T$: samples of size $m'$ from distributions $D_S$, $D_T$
  – Target error bounded by source error and unlabeled $H \Delta H$-divergence
Theory: bounding risk in target domain

- Triangle inequality for classifier error: labeling functions $f_1, f_2, f_3$,
  \[ \epsilon(f_1, f_2) \leq \epsilon(f_1, f_3) + \epsilon(f_2, f_3) \]

- Bound is now achievable
  \[
  \epsilon_T(h) \leq \epsilon_T(h^*) + \epsilon_T(h, h^*) \\
  \leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + |\epsilon_T(h, h^*) - \epsilon_S(h, h^*)| \\
  \leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + \frac{1}{2} d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T) \\
  \leq \epsilon_T(h^*) + \epsilon_S(h) + \epsilon_S(h^*) + \frac{1}{2} d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T) \\
  = \epsilon_S(h) + \frac{1}{2} d_{H \Delta H}(\mathcal{D}_S, \mathcal{D}_T) + \lambda \\
  \leq \epsilon_S(h) + \frac{1}{2} \hat{d}_{H \Delta H}(\mathcal{U}_S, \mathcal{U}_T) + 4 \sqrt{\frac{2d \log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda
  \]

- Uses H-divergence bound and VC dimension of $H \Delta H$ at most twice of H
- Any $g \in H \Delta H$ represented as linear network of depth 2 with 2 hidden units
Theory: Bound for Domain Adaptation

Domain discrepancy defined in terms of domain classifier:

\[ d_H(D_S, D_T) = 2 \sup_{h \in \mathcal{H}} | \Pr_{x \sim D_S} [h(x) = 1] - \Pr_{x \sim D_T} [h(x) = 1] | \]

(Error of the best domain classifier)

Target error bounded by source error and domain discrepancy:

\[ \epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4 \sqrt{\frac{2d \log(2m')} {m'}} + \lambda \]

Adaptability: Some model should achieve low error on both source and target