Multiple Source Domain Adaptation with Adversarial Training of Neural Networks

published: 2017-02

authors: CMU

> We propose a new generalization bound for domain adaptation when there are multiple source domains with labeled instances and one target domain with unlabeled instances.
> Compared with existing bounds, the new bound does not require expert knowledge about the target distribution.

Domain adaptation focuses on such problems by establishing knowledge transfer from a labeled source domain to an unlabeled target domain, and by exploring domain-invariant structures and representations to bridge the gap.

$$\varepsilon_T(h) \leq \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{4k}{\delta} + d\log\frac{me}{d}\right)} + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + \sqrt{\frac{2}{m}\left(\log\frac{8k}{\delta} + 2d\log\frac{me}{2d}\right)} + \lambda$$

$$= \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + \lambda + O\left(\sqrt{\frac{1}{m}\left(\log\frac{k}{\delta} + d\log\frac{me}{d}\right)}\right) \qquad (2)$$

> The first term measuers the worst case accuracy of hypothesis h on the k source domains, and the second term measures the discrepancy between the target domain and the k source domains. For domain adaptation to succeed in the multiple sources setting, we have to expect these two terms to be small: we pick our hypothesis h based on its source training errors, and it will generalize only if the discrepancy bewteen sources and target is small. The third term λ is the optimal error we can hope to achieve.

> It is also worth pointing out that these four terms appearing in the generalization bound also capture the tradeoff between using a rich hypothesis class H and a limited one as we discussed above: when using a richer hypothesis class, the first and the third terms in the bound will decrease, while the value of the second term will increase; on the other hand, choosing a limited hypothesis class can decrease the value of the second term, but we may incur additional source training errors and a large λ due to the simplicity of H.

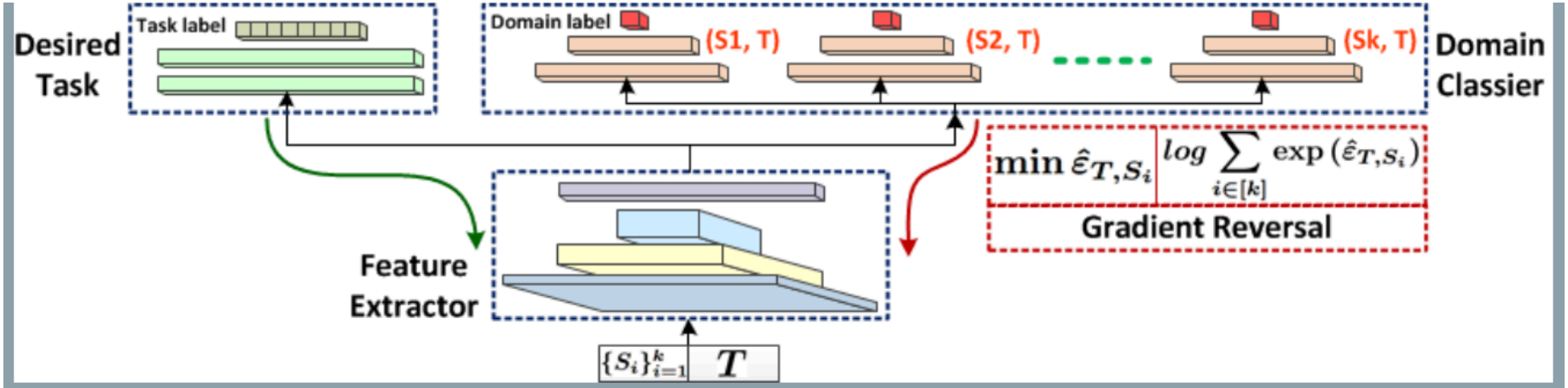

Figure 1: MDANs Network architecture. Feature extractor, domain classifier, and task learning are combined in one training process. Hard version: the source that achieves the minimum domain classification error is backpropagated with gradient reversal; Smooth version: all the domain classification risks over k source domains are combined and backpropagated adaptively with gradient reversal.

> Once we fix our hypothesis class H, the last two terms in the generalization bound will be fixed; hence we can only hope to

> minimize the bound by minimizing the first two term.

Algorithm 1 Multiple Source Domain Adaptation via Adversarial Training

for $t = 1$ to ∞ **do**
  Sample $\{S_i^{(t)}\}_{i=1}^k$ and $T^{(t)}$ from $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ and $\widehat{\mathcal{D}}_T$, each of size m
  for $i = 1$ to k **do**
    Compute $\widehat{\varepsilon}_i^{(t)} := \widehat{\varepsilon}_{S_i^{(t)}}(h) - \min_{h' \in \mathcal{H}\Delta\mathcal{H}} \widehat{\varepsilon}_{T^{(t)}, S_i^{(t)}}(h')$
    Compute $w_i^{(t)} := \exp(\widehat{\varepsilon}_i^{(t)})$
  end for
  # Hard version
  Select $i^{(t)} := \arg\max_{i \in [k]} \widehat{\varepsilon}_i^{(t)}$
  Update parameters via backpropagating gradient of $\widehat{\varepsilon}_{i^{(t)}}^{(t)}$
  # Smoothed version
  for $i = 1$ to k **do**
    Normalize $w_i^{(t)} \leftarrow w_i^{(t)} / \sum_{i' \in [k]} w_{i'}^{(t)}$
  end for
  Update parameters via backpropagating gradient of $\sum_{i \in [k]} w_i^{(t)} \widehat{\varepsilon}_i^{(t)}$
end for