Towards more Reliable Transfer Learning (Supplementary Material)

Zirui Wang and Jaime Carbonell

Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA, USA
{ziruiw,jgc}@cs.cmu.edu

A Proof of Theorem 1

First, we rewrite the definition of *ideal hypothesis* [1] between the i^{th} source and the target as:

$$h_i^* = \arg\min_{h \in \mathcal{H}} \epsilon_T(h) + \epsilon_{S_i}(h)$$

We then denote the combined risk of the i^{th} ideal hypothesis as:

$$\lambda_i = \epsilon_T(h_i^*) + \epsilon_{S_i}(h_i^*)$$

Notice that the i^{th} ideal hypothesis is a single hypothesis that performs well on *both* domains and it explicitly defines the transferability between the i^{th} source and the target. When λ_i is large, we cannot expect to transfer knowledge from the i^{th} source to the target by minimizing source error.

Similarly, for the multi-source setting, we define the ideal hypothesis on domains weighted by α and μ as:

$$h_{\alpha,\mu}^* = \arg\min_{h \in \mathcal{H}} \epsilon_T(h) + \sum_i^K (\alpha_i \mu + (1 - \alpha_i) \frac{1 - \mu}{K - 1}) \epsilon_{S_i}(h)$$

The corresponding the combined risk is then defined as:

$$\lambda_{\alpha,\mu} = \epsilon_T(h_{\alpha,\mu}^*) + \sum_i^K (\alpha_i \mu + (1 - \alpha_i) \frac{1 - \mu}{K - 1}) \epsilon_{S_i}(h_{\alpha,\mu}^*)$$

Similar to the definition above, $\lambda_{\alpha,\mu}$ defines the adaptability between weighted sources and the target, where α specifies individual source importance and μ controls the importance of inter-source relationships.

Now we can proof the bound in Theorem 1, assuming the linearity of risk:

$$\epsilon_T(\hat{h}) = \epsilon_T(\sum_i^K \alpha_i \hat{h}_i) = \sum_i^K \alpha_i \epsilon_T(\hat{h}_i) = \sum_i^K \alpha_i \Big[\mu\epsilon_T(\hat{h}_i) + \frac{1-\mu}{K-1}\sum_{j\neq i}^K \epsilon_T(\hat{h}_i)\Big]$$

$$\leq \sum_i^K \alpha_i \Bigg[\mu\Big[\epsilon_{S_i}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_i}, U_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda_i\Big] +$$

$$\frac{1-\mu}{K-1}\sum_{j\neq i}^K \Big[\epsilon_{S_j}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_j}, U_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda_j\Big]\Bigg]$$

(Theorem 1 of Blitzer)

For the combined risks of ideal hypotheses λ_i, we can rearrange and show that:

$$\sum_i^K \alpha_i \Big[\mu\lambda_i + \frac{1-\mu}{K-1}\sum_{j\neq i}\lambda_j\Big]$$

$$= \sum_i^K \alpha_i \Big[\mu(\epsilon_T(h_i^*) + \epsilon_{S_i}(h_i^*)) + \frac{1-\mu}{K-1}\sum_{j\neq i}(\epsilon_T(h_j^*) + \epsilon_{S_j}(h_j^*))\Big]$$

$$\leq \sum_i^K \alpha_i \Big[\mu(\epsilon_T(h_{\alpha,\mu}^*) + \epsilon_{S_i}(h_{\alpha,\mu}^*)) + \frac{1-\mu}{K-1}\sum_{j\neq i}(\epsilon_T(h_{\alpha,\mu}^*) + \epsilon_{S_j}(h_{\alpha,\mu}^*))\Big]$$

$$= \epsilon_T(h_{\alpha,\mu}^*) + \sum_i^K(\alpha_i\mu + (1-\alpha_i)\frac{1-\mu}{K-1})\epsilon_{S_i}(h_{\alpha,\mu}^*)$$

$$= \lambda_{\alpha,\mu}$$

Then, the risk bound becomes:

$$\leq \Bigg[\sum_i^K \alpha_i \Big[\mu\Big[\epsilon_{S_i}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_i}, U_T)\Big] + \frac{1-\mu}{K-1}\sum_{j\neq i}^K \Big[\epsilon_{S_j}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_j}, U_T)\Big]\Big]\Bigg] +$$

$$4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda_{\alpha,\mu}$$

$$\leq \Bigg[\sum_i^K \alpha_i \Big[\mu\Big[\hat{\epsilon}_{S_i}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_i}, U_T)\Big] + \frac{1-\mu}{K-1}\sum_{j\neq i}^K \Big[\hat{\epsilon}_{S_j}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_j}, U_T)\Big] +$$

$$\sqrt{\frac{\mu^2}{\beta_i} + (\frac{1-\mu}{K-1})^2\sum_{j\neq i}\frac{1}{\beta_j}}\sqrt{\frac{d\log(2m) - \log(\delta)}{2m}}\Big]\Bigg] + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda_{\alpha,\mu}$$

(Lemma 4 of Blitzer)

$$= \Bigg[\sum_i^K \alpha_i \Big[\mu\Big[\hat{\epsilon}_{S_i}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_i}, U_T)\Big] + \frac{1-\mu}{K-1}\sum_{j\neq i}^K \Big[\hat{\epsilon}_{S_j}(\hat{h}_i) + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(U_{S_j}, U_T)\Big]\Big]\Bigg] +$$

$$\Bigg(\sum_i^K \alpha_i\sqrt{\frac{\mu^2}{\beta_i} + (\frac{1-\mu}{K-1})^2\sum_{j\neq i}\frac{1}{\beta_j}}\Bigg)\sqrt{\frac{d\log(2m) - \log(\delta)}{2m}}\Bigg] + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda_{\alpha,\mu}$$

■

We also note that if $\lambda_{\alpha,\mu}$ is small, we can set the hypothesis space $\mathcal{H}$ to be the set containing all consistent hypotheses (i.e. the version space) and the weighted $\mathcal{H}\Delta\mathcal{H}$-divergence would also be small.

References

1. Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Wortman, J.: Learning bounds for domain adaptation. In: NIPS. pp. 129-136 (2008)