Fine-Tuning Graph Neural Networks via Graph Topology induced Optimal Transport

Zepeng Zhang

Conventional fine-tuning approaches can be roughly divided into two categories: (i) weight constraint, i.e., directly constraining the distance of the weights between pretrained and finetuned models. (ii) representation constraint, i.e., constrains the distance of representations produced from pretrained and finetuned models. Both types of approaches fail to take good account of the topological information.

This paper [1] proposed the graph topology induced optimal transport (GTOT) fine-tuning framework for GNN style backbones. In the GTOT-Tuning framework, the GTOT distance is used as the regularizer, which is a masked Wasserstein distance (MWD).

Definition 1 (Masked Wasserstein distance)**.** Let $\alpha = \sum_i^n \mathbf{a}_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_i^m \mathbf{b}_i \delta_{\mathbf{y}_i}$ be two discrete distributions with $\delta_{\mathbf{x}_i}$ as the Dirac function concentrated at location $\mathbf{x}_i$. The weight vectors $\sum_i^n \mathbf{a}_i = \sum_i^m \mathbf{b}_i = 1$. Given a mask matrix $\mathbf{M} \in \{0,1\}^{n\times m}$, the MWD is defined as

$$\mathbf{L}_{mw}\left(\mathbf{M}, \mathbf{a}, \mathbf{b}\right) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{M},\mathbf{a},\mathbf{b})} \left\langle \mathbf{M}\odot\mathbf{P}, \mathbf{C}\right\rangle,$$

where $\mathbf{U}\left(\mathbf{M}, \mathbf{a}, \mathbf{b}\right) = \left\{\mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \left(\mathbf{M}\odot\mathbf{P}\right)\mathbf{1}_m = \mathbf{a}, \left(\mathbf{M}\odot\mathbf{P}\right)^T\mathbf{1}_n = \mathbf{b}, \mathbf{P}\odot\left(\mathbf{1}_{n\times m} - \mathbf{M}\right) = \mathbf{0}_{n\times m}\right\}$ and $\mathbf{C} \in \mathbb{R}^{n\times m}$ is a cost matrix representing the distance between locations. With an additional entropic regularization, the MWD can be obtained in closed form.

The GTOT regularizer is a MWD with $\mathbf{C}$ calculated by cosine dissimilarity between node features, $\mathbf{M} = \mathbf{A}$ indicating the 1-hop dependence, and $\mathbf{a} = \mathbf{b} = \mathbf{1}_{|\mathcal{V}|}/\left|\mathcal{V}\right|$ being uniform distribution. Intuitively, minimizing the GTOT distance promotes the similarity of the representations produced from pretrained and finetuned models. But rather than directly promoting the representations to be identical, GTOT utilizes the local structure and actually promotes the similarity of the smoothed representations, which makes it more appropriate for the downstream tasks.

References

[1] Jiying Zhang, Xi Xiao, Long-Kai Huang, Yu Rong, and Yatao Bian. Fine-tuning graph neural networks via graph topology induced optimal transport. *arXiv preprint arXiv:2203.10453*, 2022. (document)

Preprint. Under review.