# Gradual Fine-Tuning with Graph Routing for Multi-Source Unsupervised Domain Adaptation

Yao Ma*,†, Samuel Louvan†, Zhunxuan Wang†
Amazon UK. Contact: {yaoom, slouvan, wzhunxua}@amazon.com

Third Conference on Lifelong Learning Agents

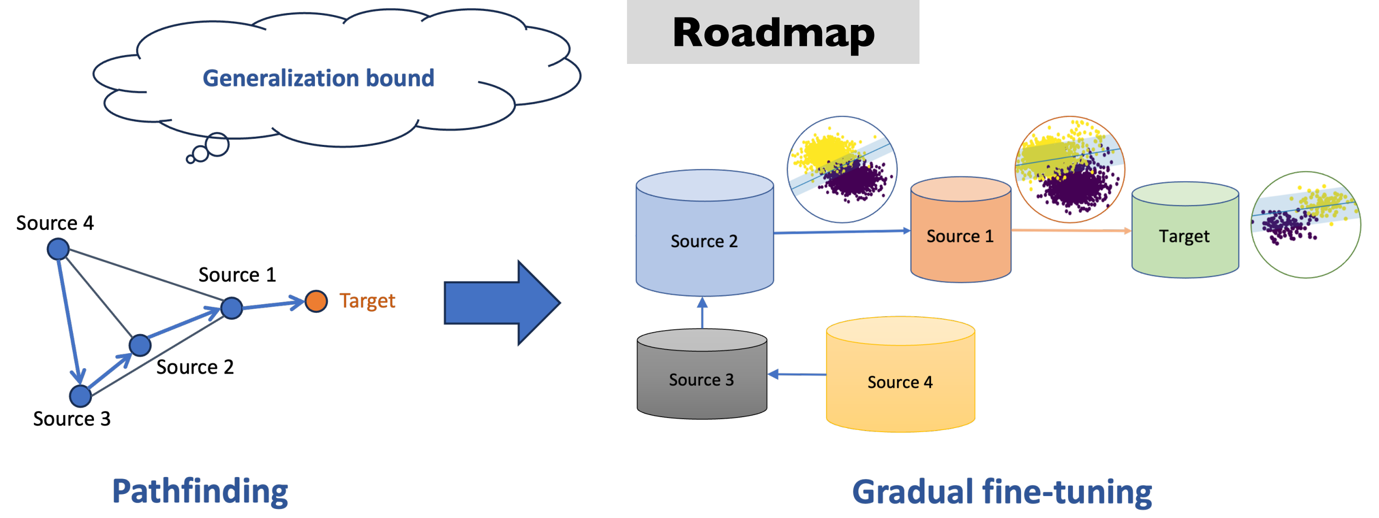
## Takeaways

Gradual fine-tuning (GFT) of models across multiple source domains to one target, represented as an undirected weighted graph.

New generalization bound for GFT along any path in graph, guiding optimal training order.

Lightweight graph-routing pathfinding strategies that minimize error bound and attain practical target accuracy efficiently.

## Roadmap




In a nutshell:

1. Derive GFT gen error bound.
2. Find path to make small bound.
3. Follow path to fine-tune model.

## Problem definition

MSUDA adapts a model to a target domain using multiple source domains without data labels in the target domain.

**Challenges**: Lack of target domain access; Costly selection; Distant sources.

**Objective**: Minimize expected target risk by training on multiple sources

$$h_T^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\{\mathbf{x},y\} \in D_T}[\mathcal{L}(h(\mathbf{x}), y)],$$

where we explore $\mathcal{H}$ by source data.

## GFT

At timestep $t$, classifier is trained on source $S_t$ with size (magnitude) $n_t$, from initialization by the previous classifier $\hat{h}_{t-1}$

$$\hat{h}_t \leftarrow \arg\min_{h \in \mathcal{H}, h^0 \leftarrow \hat{h}_{t-1}} \frac{1}{n_t} \sum_{\{\mathbf{x},y\} \in S_t} \mathcal{L}(h(\mathbf{x}), y).$$

## Generalization bound properties

- Theoretical guarantee between target errors and GFT paths.
- Distance by Wasserstein-1; Magnitude by sample size.

$$\epsilon_T \le F(\Delta\,\rho) + G(\text{mag}\,\rho),$$

$$\text{where } F' > 0 \text{ and } G' < 0.$$

**What we want**: Error bound ↘, when path length ↘ or path magnitude ↗.

## Results on MultiNLI and Sentiment Analysis

Table 1: Accuracy comparison on 5 target domains from the `MultiNLI` dataset, in mean ± std. Subscripts of average accuracies denote relative decreases to the best performance. Repeated experiments are conducted above identical set of seeds for training.

| Method | Target Domain: Fiction | Government | Telephone | Slate | Travel | Avg Acc. |
|---|---|---|---|---|---|---|
| ALL SOURCES | $76.62 \pm 0.67$ | $72.34 \pm 1.57$ | $71.94 \pm 1.37$ | $71.09 \pm 1.12$ | $72.47 \pm 2.02$ | $72.89_{(\downarrow 4.7\%)}$ |
| CLOSEST | $74.97 \pm 0.34$ | $72.88 \pm 1.19$ | $72.24 \pm 0.71$ | $73.50 \pm 1.28$ | $71.36 \pm 0.85$ | $72.99_{(\downarrow 4.6\%)}$ |
| SEAL-SHAP | $74.70 \pm 1.62$ | $75.39 \pm 0.75$ | $\mathbf{74.63} \pm 2.05$ | $73.37 \pm 0.69$ | $75.70 \pm 3.07$ | $74.75_{(\downarrow 2.3\%)}$ |
| Xu et al. (2021) | $\mathbf{78.82} \pm 1.62$ | $75.29 \pm 1.11$ | $74.97 \pm 0.59$ | $75.47 \pm 1.11$ | $73.25 \pm 2.37$ | $75.56_{(\downarrow 1.2\%)}$ |
| TGFT | $77.43 \pm 1.78$ | $\mathbf{77.19} \pm 2.13$ | $72.89 \pm 2.08$ | $74.35 \pm 1.59$ | $74.68 \pm 4.43$ | $75.30_{(\downarrow 1.6\%)}$ |
| NNGFT | $78.03 \pm 2.34$ | $76.95 \pm 2.14$ | $73.74 \pm 2.19$ | $\mathbf{77.03} \pm 6.27$ | $\mathbf{76.76} \pm 2.19$ | $\mathbf{76.50}_{(0.0\%)}$ |
| SPGFT | $76.40 \pm 1.31$ | $73.91 \pm 5.31$ | $73.05 \pm 1.69$ | $71.00 \pm 3.39$ | $73.14 \pm 2.85$ | $73.50_{(\downarrow 3.9\%)}$ |
| MSTGFT | $76.18 \pm 4.39$ | $73.91 \pm 5.31$ | $73.05 \pm 1.69$ | $71.00 \pm 3.39$ | $73.14 \pm 2.85$ | $73.45_{(\downarrow 4.0\%)}$ |

Table 3: Accuracy comparison on 4 distant domains from the multi-domain sentiment analysis dataset, in mean ± std. Subscripts of average accuracies denote relative decreases to the best performance. Repeated experiments are conducted above identical set of seeds for training.

**Efficacy in distant domains**

| Method | Target Domain: Books | Music | Electronics | Grocery | Avg Acc. |
|---|---|---|---|---|---|
| ALL SOURCES | $89.71 \pm 0.31$ | $88.83 \pm 0.67$ | $\mathbf{87.75} \pm 0.56$ | $88.66 \pm 0.53$ | $88.73_{(\downarrow 0.5\%)}$ |
| CLOSEST | $\mathbf{89.69} \pm 0.41$ | $88.98 \pm 0.22$ | $83.66 \pm 0.78$ | $88.62 \pm 0.88$ | $87.73_{(\downarrow 1.6\%)}$ |
| SEAL-SHAP | $84.85 \pm 1.80$ | $85.91 \pm 1.52$ | $88.18 \pm 0.47$ | $84.21 \pm 1.50$ | $85.79_{(\downarrow 3.8\%)}$ |
| Xu et al. (2021) | $88.87 \pm 0.85$ | $88.90 \pm 0.37$ | $87.56 \pm 0.44$ | $88.72 \pm 1.52$ | $88.51_{(\downarrow 0.7\%)}$ |
| NNGFT | $89.33 \pm 0.53$ | $\mathbf{89.85} \pm 0.32$ | $87.65 \pm 0.04$ | $\mathbf{89.85} \pm 0.82$ | $\mathbf{89.17}_{(0.0\%)}$ |

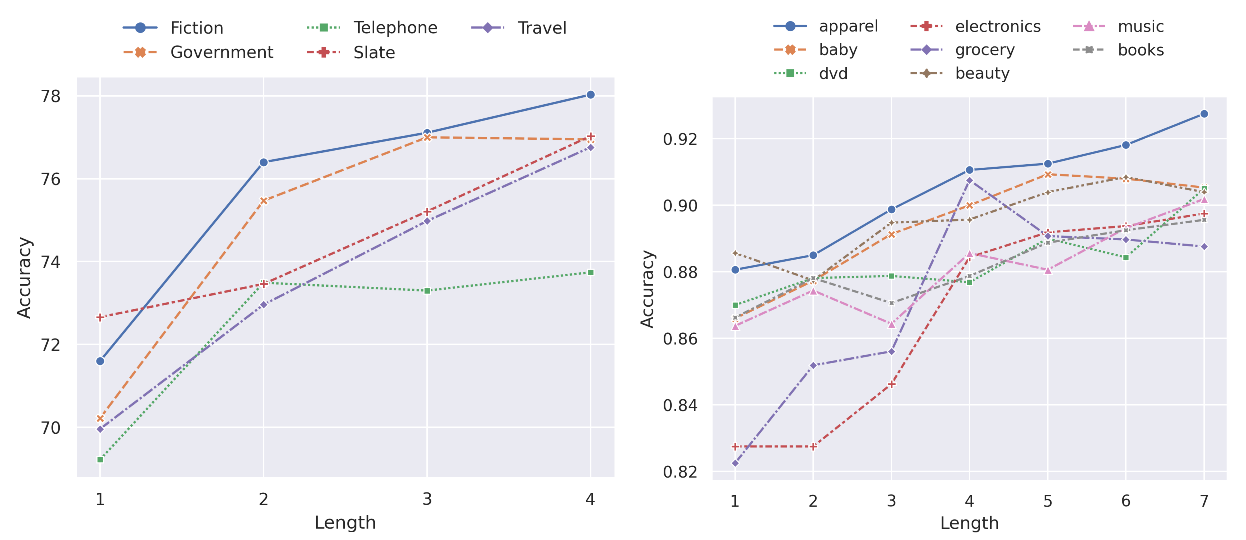
- **2.3%** and **3.9%** relative accuracy improvements over domain scoring SOTA, SEAL-Shap.
- Outperforms in-/out-domain gradual shift [Xu et al., 2021], which trains models with target data.

### Additional setup info

**Dataset**: Amazon Review for SA.
**Model**: BERT-adapted.
**Baselines**: Trained on

1. all sources combined;
2. closest source only;

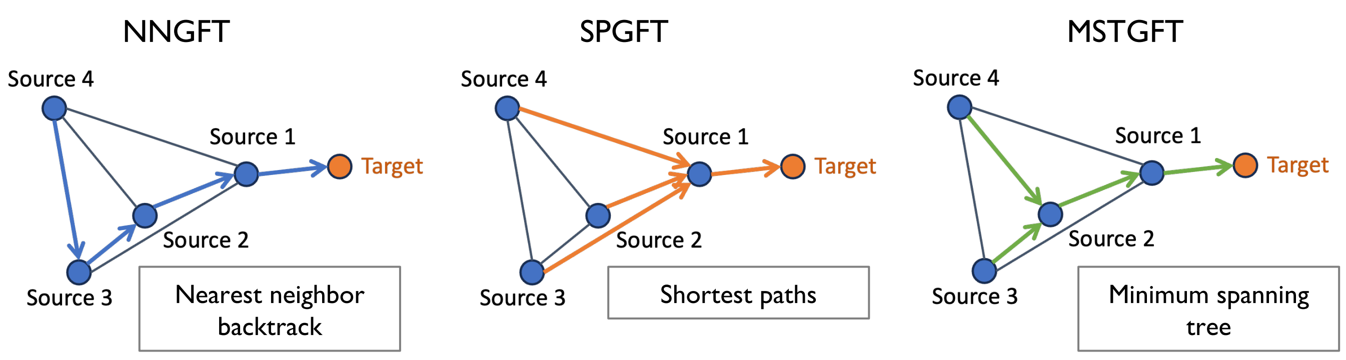
## Ablation study

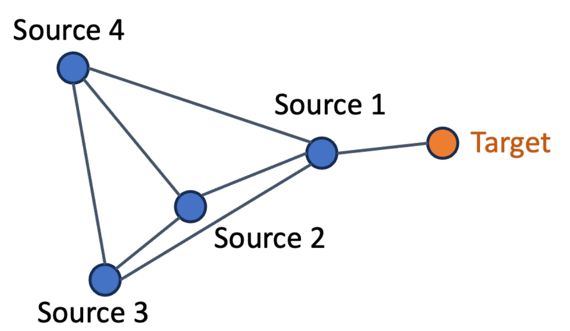



- Indicates error bound ↘ takes more advantage from path magnitude ↗.
- Justifies performance tradeoffs for SPGFT and MSTGFT, which sacrifice path magnitude for length efficiency and thereby training speed 🚀.

## Graph routing

1. Low-cost source domain selection.
2. Efficient use of distant domains.




Example Wasserstein-1 graph (edge pruned by threshold)

## Graph routing paradigm

1. For each source domain $S_i$, find the most length-efficient path to target $T$.
2. Select the one $\pi^*$ with maximum magnitude in optimal paths $P^*$.

$$\pi^* \approx \arg\max_{\pi \in P^*} \text{mag}(\pi),$$

$$\text{where } P^* = \left\{ \arg\min_{\rho \in P_G(S_i, T)} \Delta(\rho),\ i \in [K] \right\}.$$

## Future work

1. Gap between UDA generalization error and graph theory.
2. Source combining + pathfinding instead of source-by-source.
3. Scalability to larger domain graphs.
4. Exact / Stronger path optimality (not just error bounded).
5. More expressive distance measure.

Join the climb: Scan to read!

* Corresponding author
† Equal contribution