Approximate Clustering for Extracting Task Relationships in Multi-Instruction Tuning

Dongyue Li, Jinhong Yu, Hongyang R. Zhang
Email addresses: {li.dongyu, yu.jinh, ho.zhang}@northeastern.edu

Problem Statement

Multitask learning problems in fine-tuning language models:

- Multitask instruction Fine-tuning: Fine-tuning a language model on multiple NLP tasks
- Multi-instruction fine-tuning: Fine-tuning a model on a mix of instructions.
- In-context learning: Learning a model to solve different function classes with in-context examples.

Task Grouping. Given n tasks, the goal is to partition the n tasks into k subsets such that each subset is the best to be trained together.

- **Task affinity score** $T_{u,v}$ denotes the transfer effect from u to v.
- **Density of affinity scores in a subset** characterizes the extent of positive transfers within a subset S:

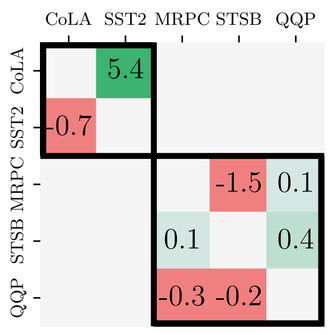
$$d_S = \sum_{u,v \in S} \frac{T_{u,v}}{|S|}$$

Task grouping as a clustering problem that aims to maximize the average density of all clusters:

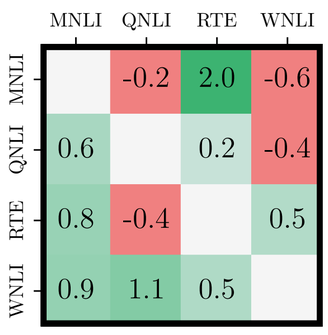
$$\sum_{i=1}^{k} d_{C_i} = \sum_{i=1}^{k} \sum_{u,v \in C_i} \frac{T_{u,v}}{|C_i|} = \sum_{i=1}^{k} \frac{v_i^\top T v_i}{v_i^\top v_i}.$$

$C_1, \ldots, C_k$ denote a partition of the n tasks. $v_1, \ldots, v_k$ denote 0-1 vectors indicating whether each task is in the cluster or not.

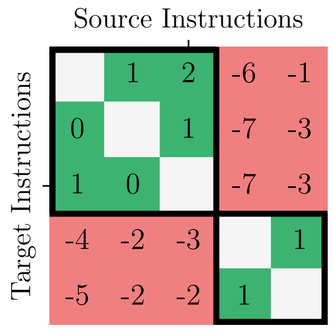
Examples of negative interference between tasks. For each entry, we combine one target task with another task and report the performance difference of multitask minus single-task learning.



(a) Single-Sentence & Similarity and Paraphrase Tasks from GLUE



(b) Four Natural Language Inference Tasks from GLUE



(c) Five different instructions from PromptSource on the RTE dataset

Approximate Task Grouping

Clustering through semidefinite programming relaxations. Integer program is computationally challenging to solve:

$$\max \left\{ \langle T, \sum_{j=1}^{k} \frac{v_j v_j^\top}{v_j^\top v_j} \rangle : Ve = e, \sum_{i=1}^{n} V_{i,j} \geq 1 \text{ for } 1 \leq j \leq k, V \in [0,1]^{n \times k} \right\}$$

Relax the integer program to a semidefinite program (SDP):

$$\max \left\{ \langle T, X \rangle : Xe = e, \sout{r(X) = k}, \mathrm{Tr}[X] = k, X \geq 0, X \geq 0, X \in \mathbb{R}^{n \times n} \right\}$$

Round the solution by thresholding via $\frac{c}{n}$ with $c > 1$.

Efficiently estimating task affinities.

- *Higher-Order Task Affinity:* Estimate task affinity score from subsets of more than two tasks.

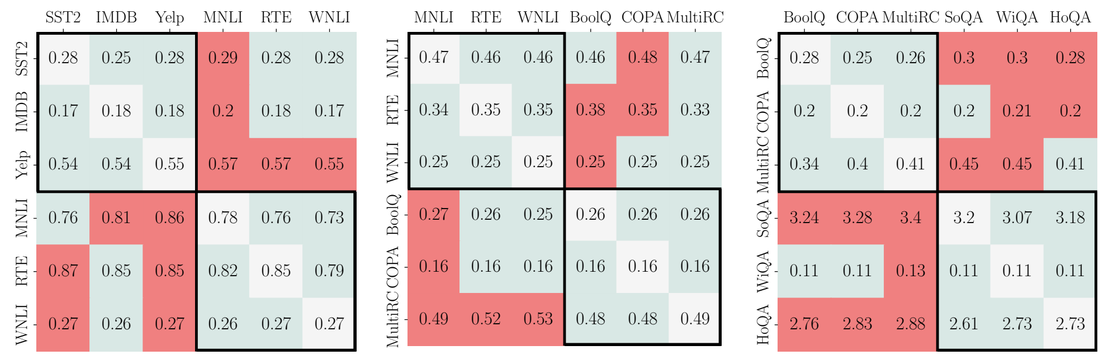
$$T_{i,j} = \frac{1}{n_{i,j}} \sum_{1 \leq k \leq n : \{i,j\} \subseteq S} f_i(S_k), \text{ for all } 1 \leq i, j \leq n$$

- *Adaptive Sampling:* Iteratively estimate affinity scores for small batches of tasks. In each iteration, pick one cluster from current clusters to estimate task affinity scores between it and new tasks.

Evaluations of Task Grouping

Task Grouping for Multitask Instruction Fine-Tuning on 19 NLP datasets under 6 categories: sentiment analysis, NLI, multiple-choice QA, open-domain QA, coreference, and summarization tasks.

- We measure the pairwise transfers between each pair of tasks and select the subsets whose ratio of positive effects > 90%.
- Our approach correctly identifies the groups of all cases (matches the exhaustive search.). Spectral and Lloyd's clustering identify the group structures in 16 and 4 out of the 57 cases.

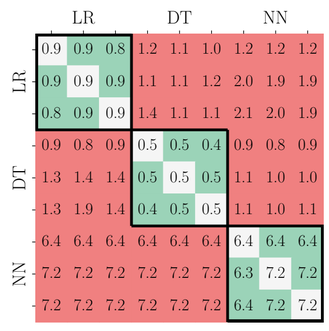


Multi-Instruction Tuning. Our approach, clustering 100 instructions into 10 groups, improves over the baseline methods by 3.3%.

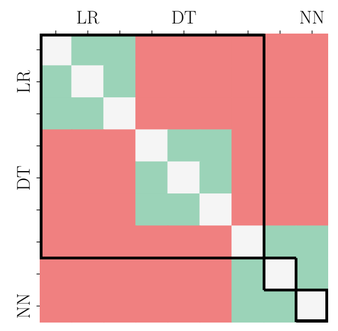
Dataset Metric	RTE	WiC Accuracy	BoolQ	E2E NLG ROGUE-1	Web NLG ROGUE-1
Multi-Instruction Tuning	75.09	66.44	78.16	71.46	80.80
Prefix Tuning	72.74	62.29	76.19	70.23	78.69
Prompt Tuning	73.12	62.88	75.51	70.72	77.42
Grouping by Spectral Clustering	73.18	65.09	75.71	71.91	81.27
Grouping by Lloyd's Clustering	73.58	64.61	75.61	71.26	80.41
Our Approach	**80.96**	**69.89**	**81.76**	**73.03**	**82.95**

In-Context Learning. Training transformers to in-context learn three types of functions (Garg et al. '22): linear regression (LR), decision trees (DT), and two-layer neural networks (NN). We define 3 function classes with different distributions of each function type.

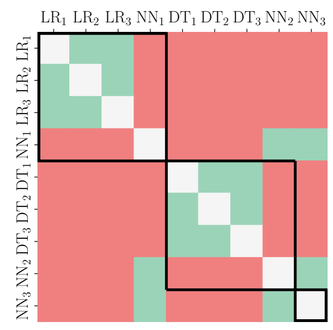
- Transformers trained on different function classes perform worse than being trained on a single function class.



SDP relaxation

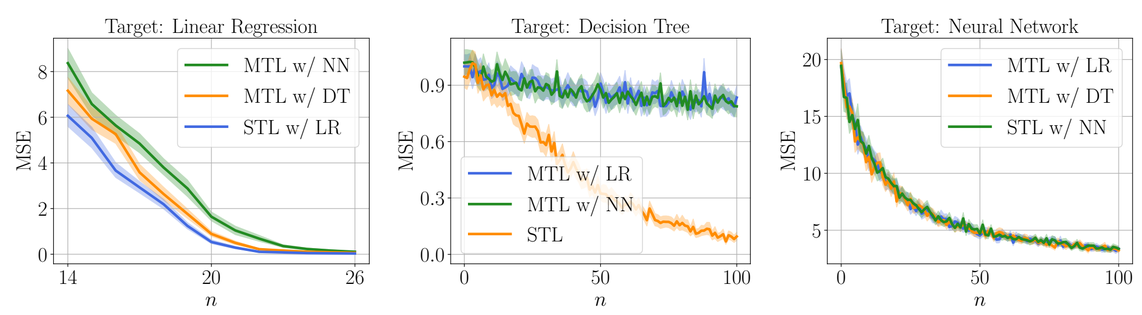


Spectral clustering



Lloyd's clustering

In-context transferability. A transformer trained on NNs with LR or DT compares comparably to a transformer trained only on NNs. Learning DT (or LR) with others will significantly degrade MSE.



Conclusion

- An approximate clustering algorithm to group tasks in language model fine-tuning
- An evaluation benchmark of 67 cases for task grouping in three scenarios of language model fine-tuning.
- Our algorithm correctly identifies underlying group structures and improves multi-instruction tuning.

Lab website: virtuosoresearch.github.io
Paper: openreview.net/forum?id=CZJOOFgXZj
Code: anonymous.4open.science/r/AdaGroup4InstructionTuning