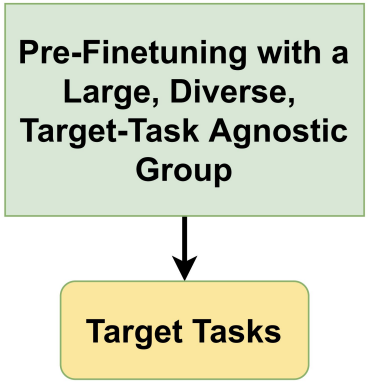
Exploring the Role of Task Transferability in Large-Scale Multi-Task Learning

Vishakh Padmakumar[1], Leonard Lausen[2], Miguel Ballesteros[3], Sheng Zha[2], He He[1,2], George Karypis[2]

[1]New York University, [2]AWS AI, [3]AWS AI Labs

To Scale or Not To Scale, That Is The Question

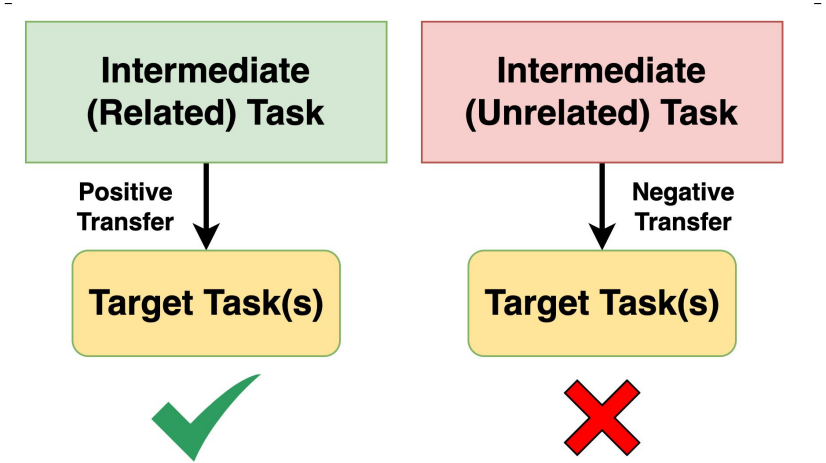
- Multi-task pre-finetuning on a ***sufficiently large, diverse*** set of tasks is an effective ***task-agnostic*** second-stage of model pre-training[1]



[1] Aghajanyan, Armen, et al. "Muppet: Massive multi-task representations with pre-finetuning." *arXiv preprint arXiv:2101.11038* (2021).

To Scale or Not To Scale, That Is The Question

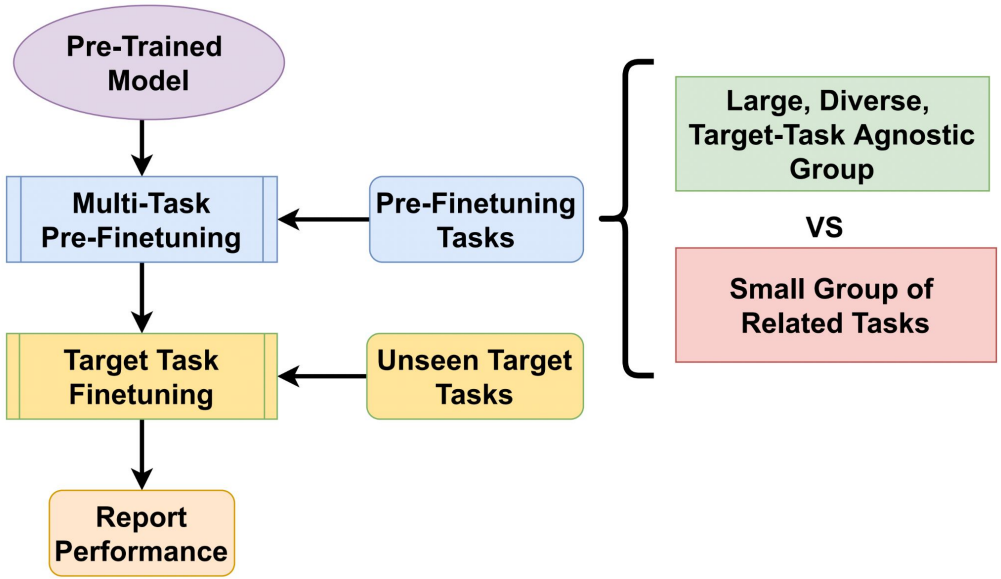
- Work on transferability has shown that the ***choice of intermediate task*** significantly impacts downstream task performance[2,3]



[2] Vu, Tu, et al. "Exploring and predicting transferability across NLP tasks." *arXiv preprint arXiv:2005.00770* (2020).
[3] Pruksachatkun, Yada, et al. "Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?." *arXiv preprint arXiv:2005.00628* (2020).

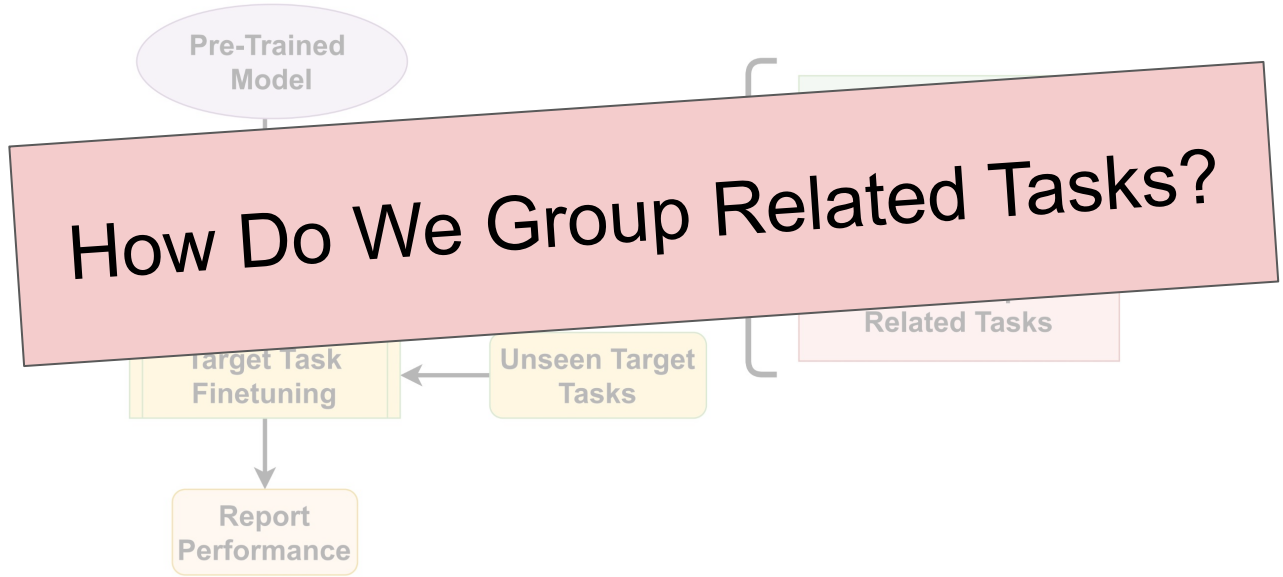
Research Question

- We aim to study how the ***choice of pre-finetuning tasks*** and the ***size of the multi-task step*** affects target task performance.

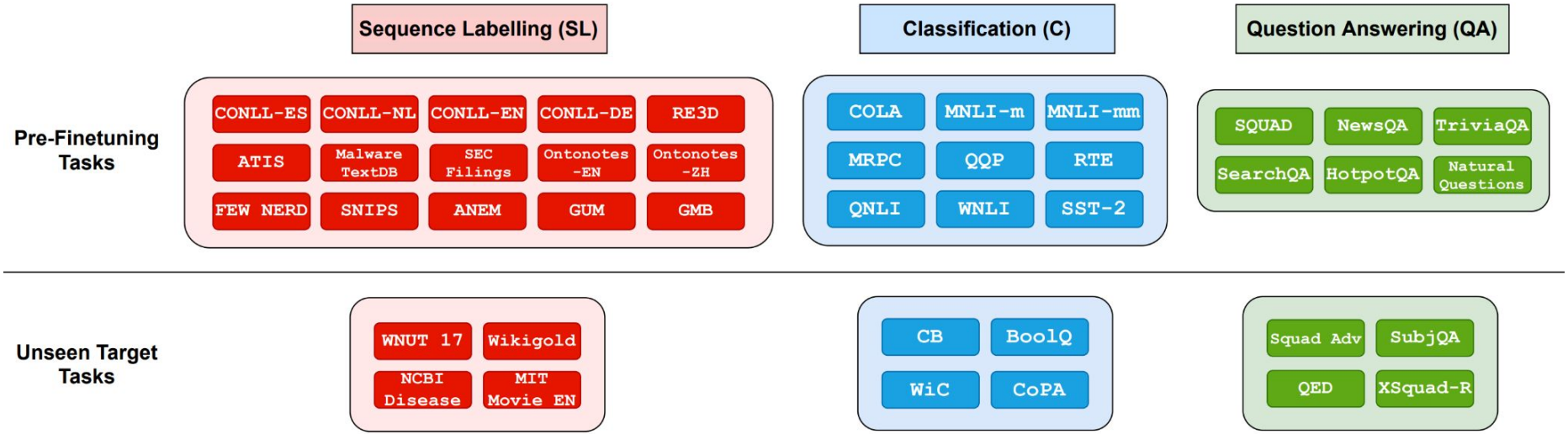


Research Question

- We aim to study how the ***choice of pre-finetuning tasks*** and the ***size of the multi-task step*** affects target task performance.



How Do We Group Related Tasks?



Experiments

Run pre-finetuning on each combination of task groups

	Baseline	Only-SL	Only-C	Only-QA	SL+C	SL+QA	QA+C	SL+C+QA
Unseen SL	80.134	**80.844**	72.370	78.693	79.453	80.165	77.378	80.750
Unseen C	68.109	67.406	71.404	63.21	70.422	70.068	70.067	**73.021**
Unseen QA	56.692	45.174	61.120	75.252	57.568	75.460	75.035	**75.678**
Average	68.312	64.475	68.298	72.385	69.147	75.564	74.160	**76.483**

Experiments

Report results on all unseen tasks, averaged over the task groups

	Baseline	Only-SL	Only-C	Only-QA	SL+C	SL+QA	QA+C	SL+C+QA
Unseen SL	80.134	**80.844**	72.370	78.693	79.453	80.165	77.378	80.750
Unseen C	68.109	67.406	71.404	63.21	70.422	70.068	70.067	**73.021**
Unseen QA	56.692	45.174	61.120	75.252	57.568	75.460	75.035	**75.678**
Average	68.312	64.475	68.298	72.385	69.147	75.564	74.160	**76.483**

Effect of Multi-Task Scaling

On average, a large-scale task-agnostic multi-task step improves downstream performance

	Baseline	Only-SL	Only-C	Only-QA	SL+C	SL+QA	QA+C	SL+C+QA
Unseen SL	80.134	**80.844**	72.370	78.693	79.453	80.165	77.378	80.750
Unseen C	68.109	67.406	71.404	63.21	70.422	70.068	70.067	**73.021**
Unseen QA	56.692	45.174	61.120	75.252	57.568	75.460	75.035	**75.678**
Average	68.312	64.475	68.298	72.385	69.147	75.564	74.160	**76.483**

Effect of Transferability

Target task performance on unseen tasks is improved when we pre-finetuning on related tasks from the same group

	Baseline	Only-SL	Only-C	Only-QA	SL+C	SL+QA	QA+C	SL+C+QA
Unseen SL	80.134	**80.844**	72.370	78.693	79.453	80.165	77.378	80.750
Unseen C	68.109	67.406	71.404	63.21	70.422	70.068	70.067	**73.021**
Unseen QA	56.692	45.174	61.120	75.252	57.568	75.460	75.035	**75.678**
Average	68.312	64.475	68.298	72.385	69.147	75.564	74.160	**76.483**

Interplay of Transferability and Multi-Task Scaling

Target task performance on pre-finetuning with a small group of related tasks is on-par with the large-scale multi-task setup

	Baseline	Only-SL	Only-C	Only-QA	SL+C	SL+QA	QA+C	SL+C+QA
Unseen SL	80.134	**80.844**	72.370	78.693	79.453	80.165	77.378	80.750
Unseen C	68.109	67.406	71.404	63.21	70.422	70.068	70.067	**73.021**
Unseen QA	56.692	45.174	61.120	75.252	57.568	75.460	75.035	**75.678**
Average	68.312	64.475	68.298	72.385	69.147	75.564	74.160	**76.483**

Interplay Between Task Groups

It's hard to predict the interplay between tasks, so selecting an optimum subset of tasks is challenging

	Baseline	Only-SL	Only-C	Only-QA	SL+C	SL+QA	QA+C	SL+C+QA
Unseen SL	80.134	**80.844**	72.370	78.693	79.453	80.165	77.378	80.750
Unseen C	68.109	67.406	71.404	63.21	70.422	70.068	70.067	**73.021**
Unseen QA	56.692	45.174	61.120	75.252	57.568	75.460	75.035	**75.678**
Average	68.312	64.475	68.298	72.385	69.147	75.564	74.160	**76.483**

Similar findings were reported in Aribandi, Vamsi, et al. "ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning." *arXiv preprint arXiv:2111.10952* (2021).

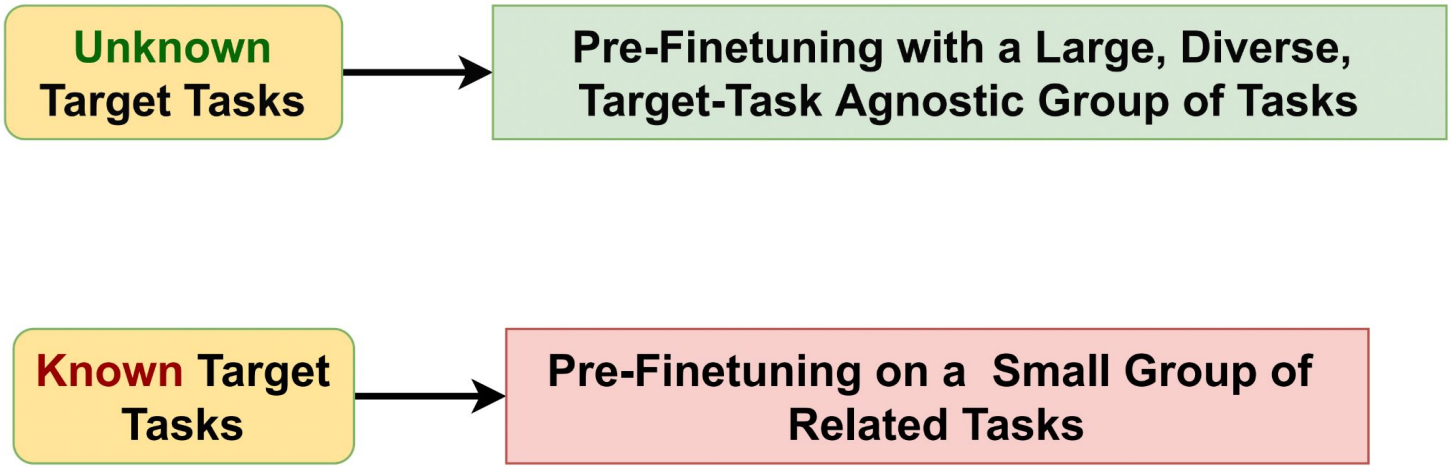
Takeaways

- When the target tasks are **unknown**, **multi-task scaling** provides an effective **general purpose model**

Takeaways

- If the goal is to improve performance on a **specific target task(s)** then a **smaller set of related tasks** is an effective, **computationally cheaper alternative**

Takeaways


Unknown Target Tasks
Pre-Finetuning with a Large, Diverse, Target-Task Agnostic Group of Tasks
Known Target Tasks
Pre-Finetuning on a Small Group of Related Tasks

For more details, stop by
Poster Session 2 on 7/11 at 2:30pm :)