Eliciting and Understanding Cross-task Skills with Task-level Mixture-of-Experts

Qinyuan Ye, Juan Zha, Xiang Ren {qinyuany, juanzha, xiangren}@usc.edu

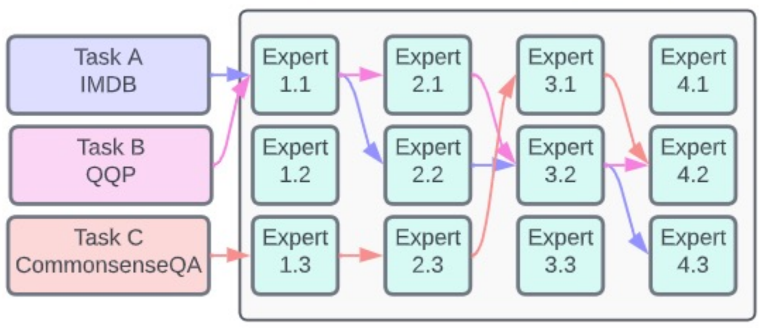
TL;DR

- We train task-level mixture-of-experts models to multi-task on diverse NLP tasks.
- They are better at generalizing to unseen tasks in few-shot and zero-shot setting.
- Learned routes and experts partly align with human categorization of NLP tasks.

1. Motivation

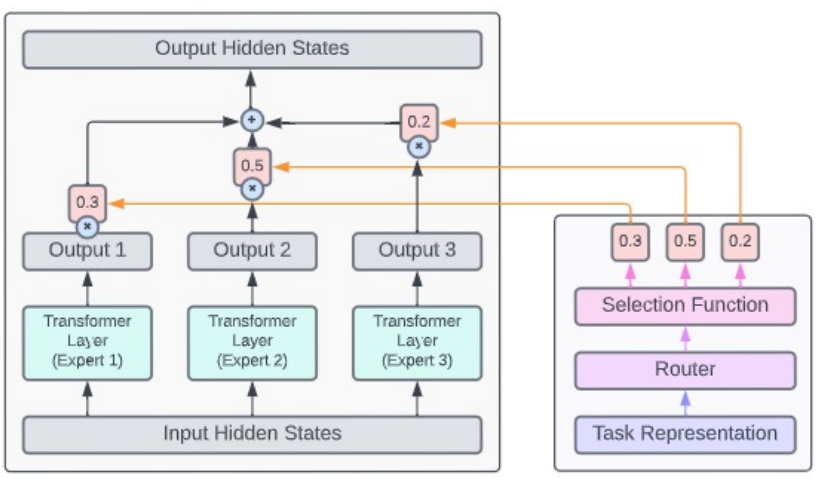
- Training transformer models to multi-task is beneficial. However, the potential of these models may be limited as they use the exact ***same*** set of parameters for very ***different*** tasks.
- Humans, on the other hand, develop skill sets and accumulate knowledge during learning, and reuse only the necessary skills when facing a new task.
- ***What if…*** we train a multi-task model that explicitly emulate skill and knowledge sharing?

→ **Task-level Mixture-of-Experts**



Does this help multi-task learning?
What is learned by each expert?
Does this improve generalization to new tasks?

2. Task-level Mixture-of-Experts



Mixture-of-Experts Transformer Layer   Router

Router: Selects and decides which experts to use for each task at each layer, based on the (trainable) task representations.

Experts: We copied the n transformer blocks in the original model for m times, resulting in $m*n$ transformer blocks in total. We assume that each transformer block is acting as an expert in that layer. (n=12, m=3)

3. Data

Setting	Dataset	# Seen/Unseen Tasks
Few-shot	CrossFit (Ye et al., 2021)	120/18
Zero-shot	Public Pool of Prompts, P3 (Sanh et al., 2021; Bach et al., 2021)	35/10

4. Comparing Different Design Choices

Things that matter

Selection	Batching	Two-speed LR	Two-stage Training
Softmax	Heterogenous	Yes	Yes
Gumbel Softmax	Homogeneous	No	No
Gumbel Softmax w/ Straight Through			

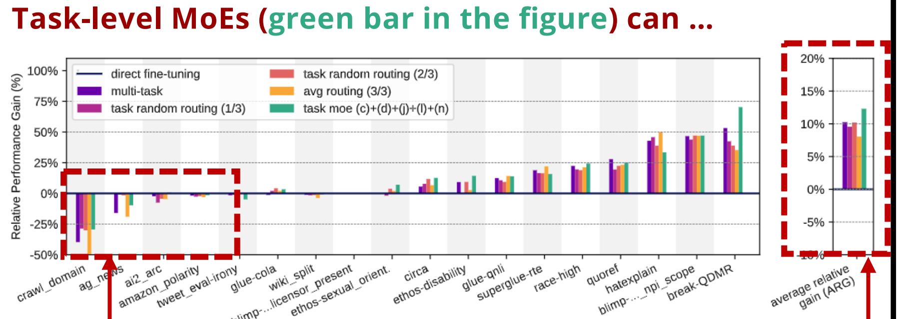
Things that don't matter much

Router	Task Representation	Freeze Task Representation
MLP	Random	Yes
LSTM	Text Embedding	No
Transformer	Fisher Information Task Embedding	

5. Few-shot Adaptation to Unseen Tasks

Comparing with… BART-Base trained with **(1)** vanilla multi-tasking **(2)** random task routing (1/3, 2/3); **(3)** avg routing (3/3)

Task-level MoEs (green bar in the figure) can …



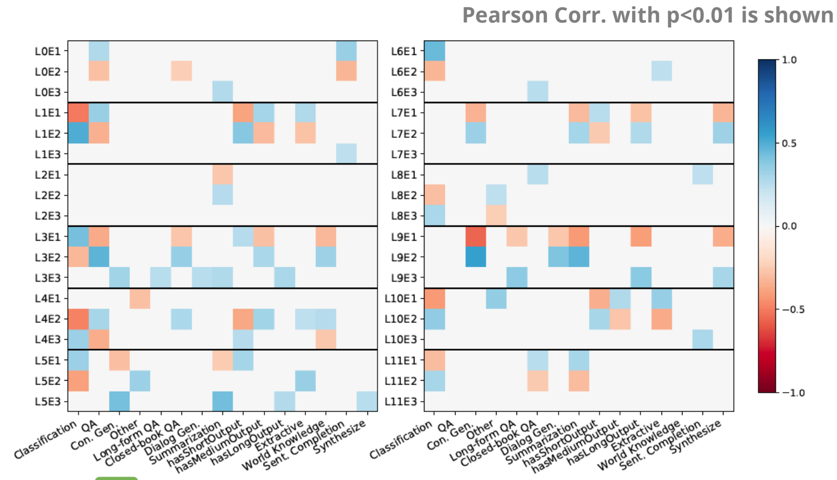
Avoid negative transfer   Generalize better to unseen tasks

Check-out zero-shot performance on P3 dataset in our paper!

6. Understanding the Learned Routes

Correlation between learned routes and hand features

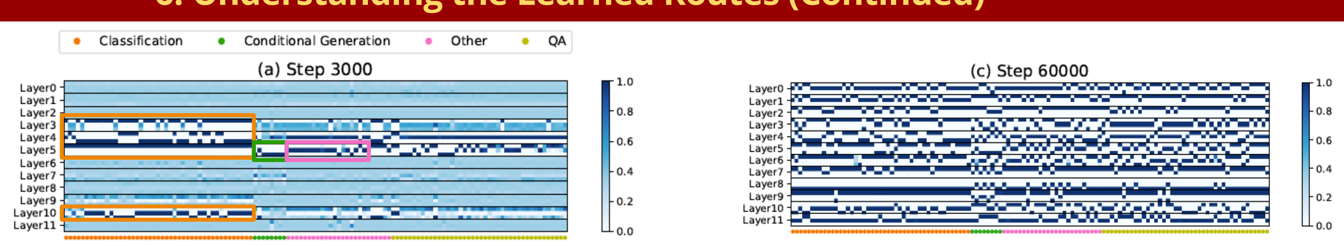
Pearson Corr. with p<0.01 is shown



Verified with expert disabling experiments

6. Understanding the Learned Routes (Continued)

Learning dynamics of routes



Developing patterns early on → Becoming more fine-grained and discrete gradually