Continual Task Allocation in Meta-Policy Network via Sparse Prompting

Yijun Yang[1,2], Tianyi Zhou[3], Jing Jiang[2], Guodong Long[2], Yuhui Shi[1]

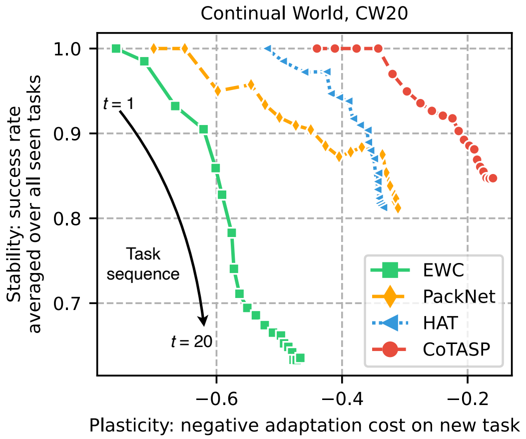
TL;DR: We propose **CoTASP**, a continual RL method, which learns **dictionaries** to extract a "**winning ticket**" (sparse subnetwork) for each task so only semantically similar tasks share more knowledge while avoiding negative task interference. We used a pretrained language model to capture the task similarity and understand task instruction, thus enabling zero-shot generalization to new tasks.

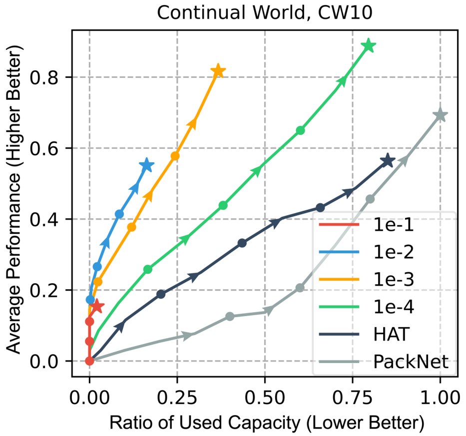
Plasticity-stability trade-off is a primary challenge in continual reinforcement learning (RL):

a. The policy needs to retain the knowledge shared across different tasks in history (**stability**)
b. While quickly adapting to new tasks without interference from previous irrelevant tasks (**plasticity**).

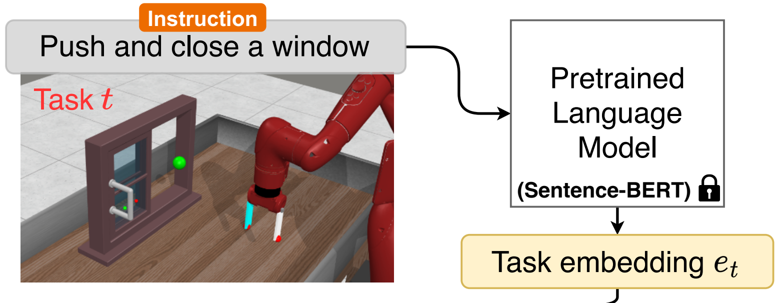
CoTASP achieves a better trade-off by learning a generalizable meta-policy network that

- Sparsely represents past tasks' knowledge, allowing more efficient usage of model capacity (**better stability**),
- And selectively reuses the knowledge from previous tasks for quick adaptation to new tasks (**better plasticity**).



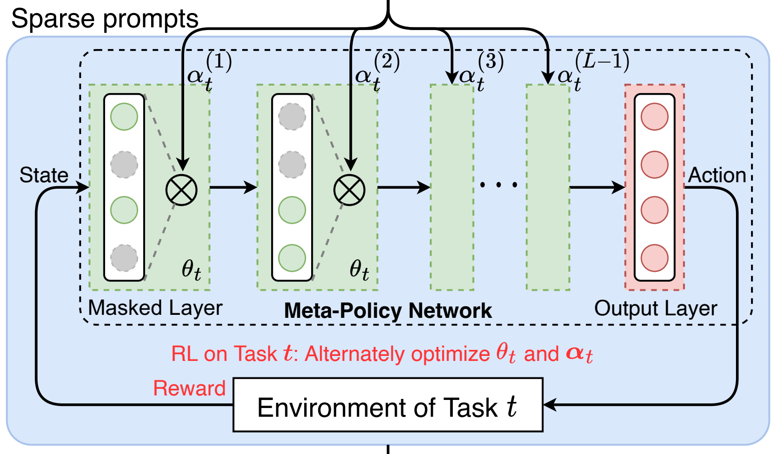


How does CoTASP work?



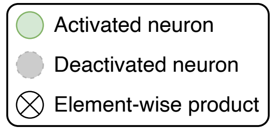
Sparse coding of e_t

$$\alpha_t^{(l)} = \arg\min_\alpha \frac{1}{2}\|e_t - \mathbf{D}_{t-1}^{(l)}\alpha\|_2^2 + \lambda\|\alpha\|_1 \text{ for layer } l = 1, \dots, L-1$$



Dictionary learning

$$\mathbf{D}_t^{(l)} = \arg\min_{\mathbf{D}} \frac{1}{2}\sum_{i=1}^{t}\|e_i - \mathbf{D}\alpha_i^{*(l)}\|_2^2,$$
$$\text{s.t. } \mathbf{D}[j]^{\mathrm{T}}\mathbf{D}[j] \le c\ \forall j, \text{ for } l = 1, \dots, L-1$$



- Relevant tasks extract similar sub-networks from the embedding of task instructions via **sparse coding**;
- **Dictionaries are learned** to align optimized masks $\boldsymbol{\alpha}^*$ with tasks' embedding, thereby capturing semantic correlations.
- Given dictionaries learned on previous tasks, **new task adaptation** reduces to highly efficient sparse prompting and sub-network finetuning.

Experimental Results on Continual World

Evaluation (mean±std of 3 metrics over 5 random seeds) on Continual World.
P: Average success rate, F: Forgetting metric, G: Generalization metric

Benchmarks		CW 10			CW 20		
Metrics		P (↑)	F (↓)	G (↓)	P (↑)	F (↓)	G (↓)
Reg	L2	0.44 ±0.12	0.00 ±0.06	0.51 ±0.07	0.52 ±0.07	-0.10 ±0.05	0.58 ±0.06
	EWC	0.64 ±0.14	0.02 ±0.05	0.34 ±0.04	0.60 ±0.07	0.02 ±0.03	0.39 ±0.06
	MAS	0.60 ±0.14	-0.06 ±0.04	0.44 ±0.07	0.48 ±0.06	0.02 ±0.02	0.49 ±0.03
	VCL	0.48 ±0.10	-0.02 ±0.06	0.43 ±0.06	0.50 ±0.11	-0.04 ±0.08	0.52 ±0.06
	Finetuning	0.12 ±0.04	0.70 ±0.04	0.25 ±0.06	0.05 ±0.00	0.72 ±0.03	0.30 ±0.05
Struc	PackNet	0.80 ±0.09	0.00 ±0.00	0.28 ±0.07	0.78 ±0.07	0.00 ±0.00	0.32 ±0.04
	HAT	0.68 ±0.12	0.00 ±0.00	0.44 ±0.07	0.67 ±0.08	0.00 ±0.00	0.46 ±0.04
	TaDeLL	0.75 ±0.04	0.00 ±0.00	0.68 ±0.01	0.66 ±0.03	0.01 ±0.02	0.67 ±0.01
Reh	Reservoir	0.32 ±0.12	0.04 ±0.05	0.79 ±0.02	0.08 ±0.09	0.14 ±0.05	0.87 ±0.01
	A-GEM	0.14 ±0.05	0.68 ±0.04	0.23 ±0.02	0.08 ±0.02	0.72 ±0.07	0.29 ±0.04
	ClonEx-SAC*	0.86	0.02	–	0.87	0.02	–
MT	MTL	0.52 ±0.10	–	–	0.50 ±0.11	–	–
	MTL+PopArt	0.70 ±0.14	–	–	0.66 ±0.17	–	–
	CoTASP (ours)	0.92 ±0.04	0.00 ±0.00	0.24 ±0.03	0.88 ±0.02	0.00 ±0.00	0.27 ±0.03

a) New task adaptation. **b)** Convergence of dictionary learning. **c)** subnetwork mask similarity