DeepMind

A Combinatorial Perspective on Transfer Learning

Jianan Wang, Eren Sezener, David Budden, Marcus Hutter, Joel Veness

Introduction

In this work we study how the learning of modular solutions can allow for effective generalization to both unseen and potentially differently distributed data.

Our main postulate is that the combination of task segmentation, modular learning and memory-based ensembling can give rise to **generalization on an exponentially growing number of unseen tasks.**

We provide a concrete instantiation of this idea and demonstrate that this system exhibits a number of desirable continual learning properties: robustness to catastrophic forgetting, no negative transfer and increasing levels of positive transfer as more tasks are seen. We show competitive performance against both offline and online methods on standard continual learning benchmarks.

Algorithm Setup

Node-level modular learning algorithm:

Gated Geometric Mixer (GGM)

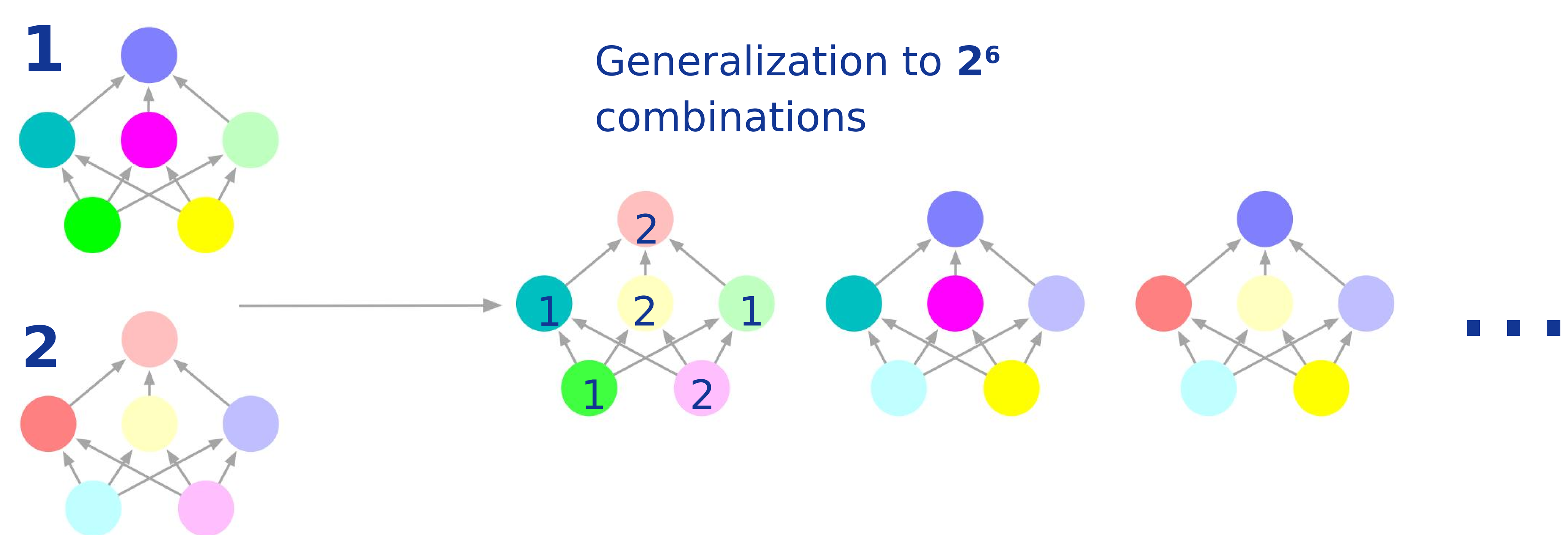
- Well studied ensemble technique for combining probabilistic forecasts
- Basic building block for Gated Linear Networks (GLNs) [1, 2]

Automatic task segmentation and local ensembling:

Forget-Me-Not (FMN) Process [3]

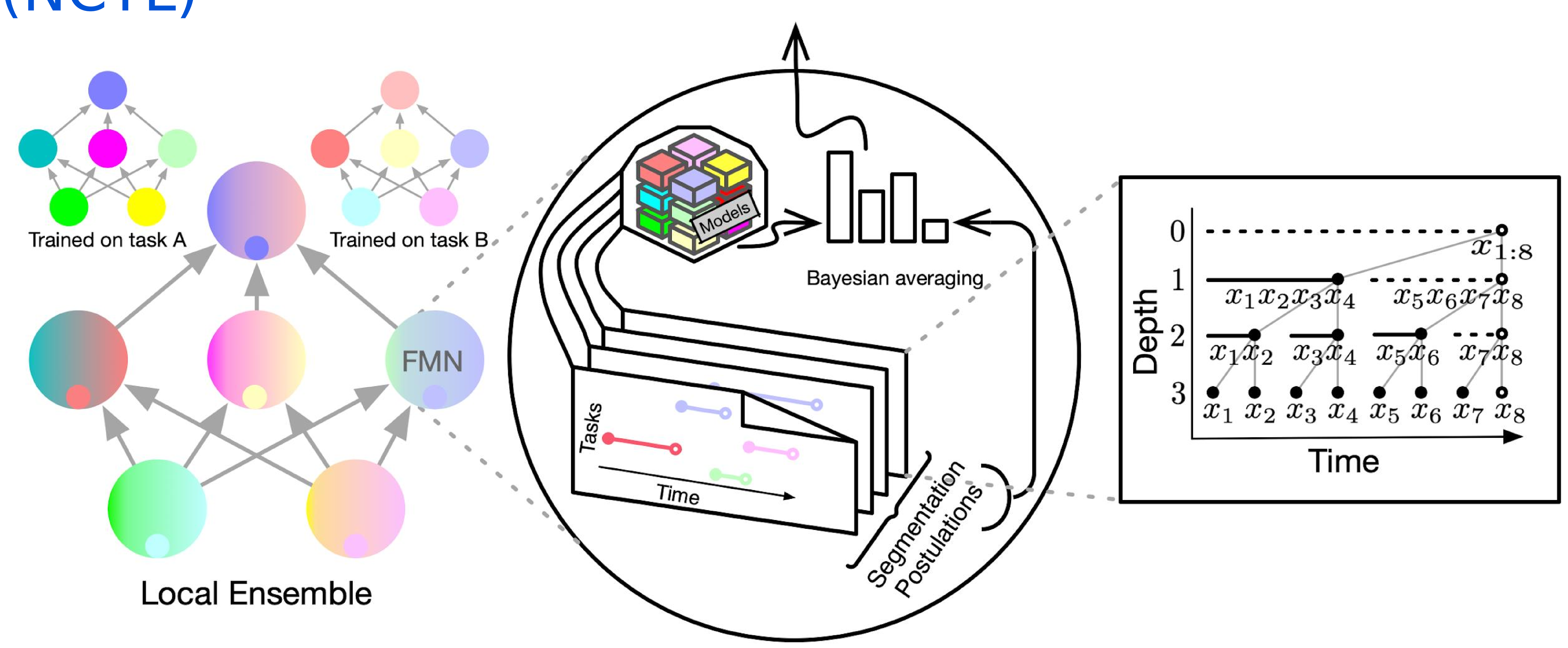
- Efficient online Bayesian changepoint detection/task identification/reuse of previous learnt solutions

Thought Experiment: the Power of Modularity



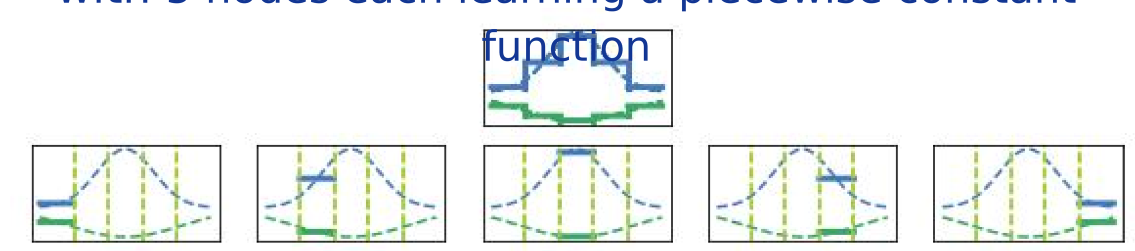
A network with m nodes trained on k tasks will generalize to k^m pseudo-tasks !

Neural combinatorial transfer learning (NCTL)

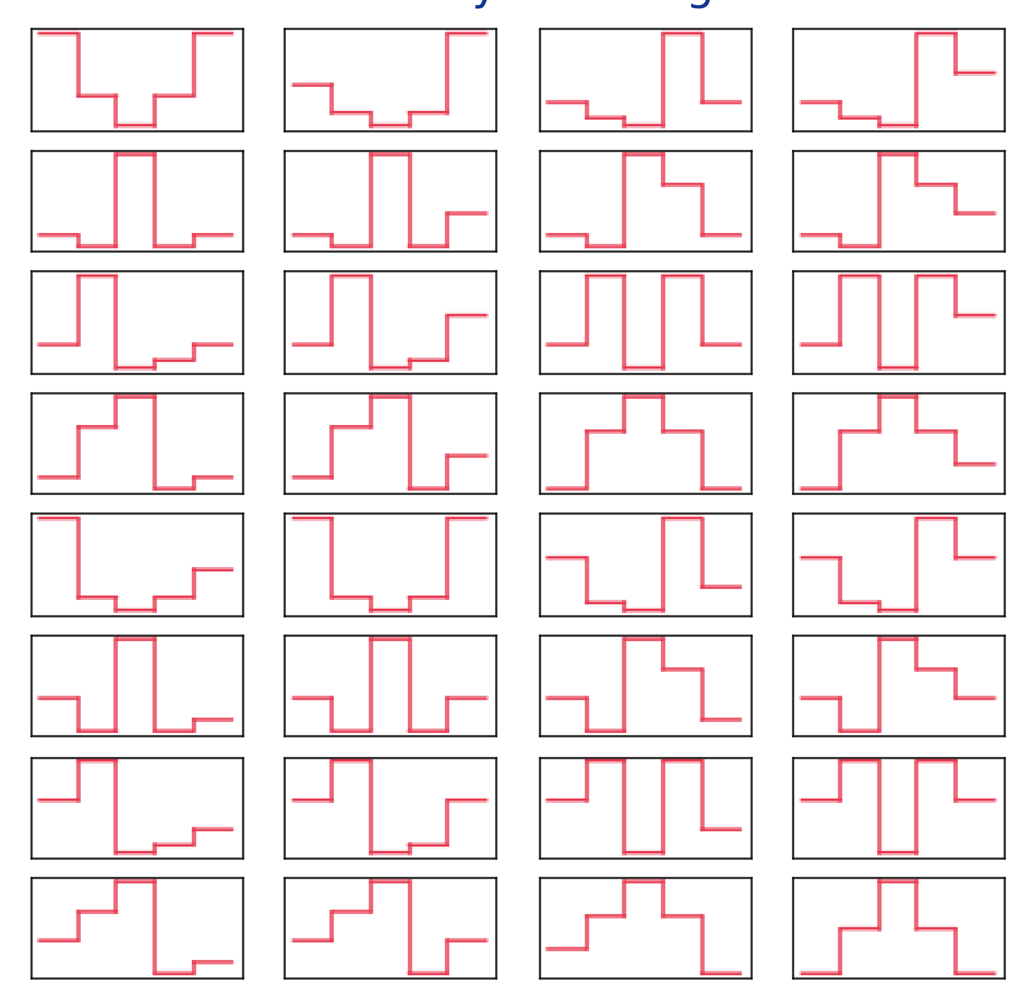


Toy Example: the Power of Modularity

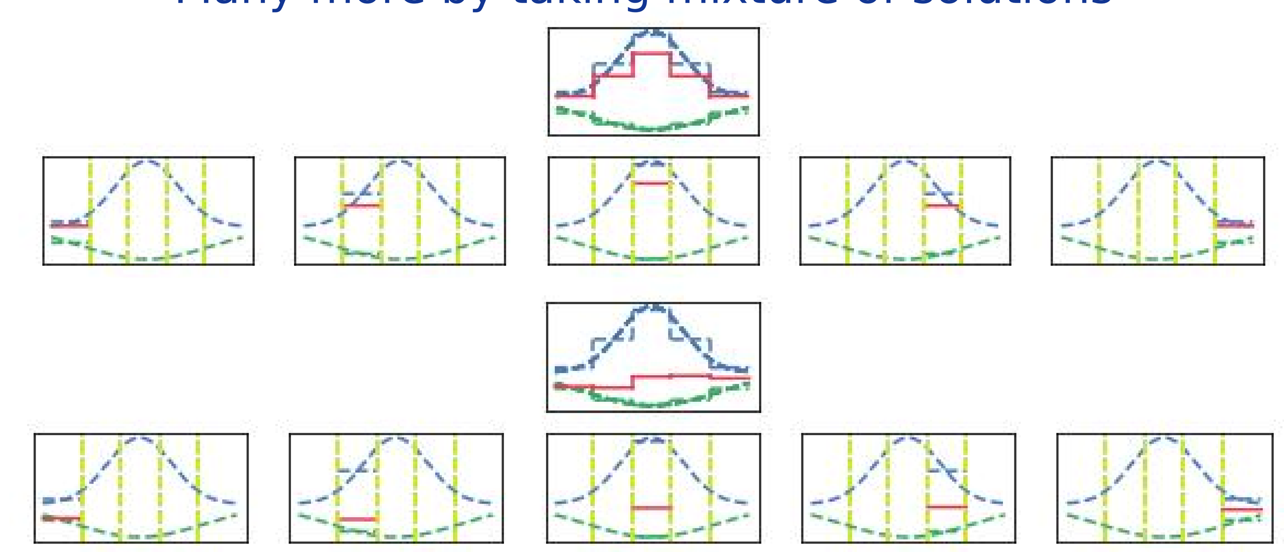
Approximating two Gaussian functions with 5 nodes each learning a piecewise constant function



32 immediate initializations by choosing one solution at each n



Many more by taking mixture of solutions



...

Experimental Results

We evaluated NCTL on diagnostic and benchmark tasks.

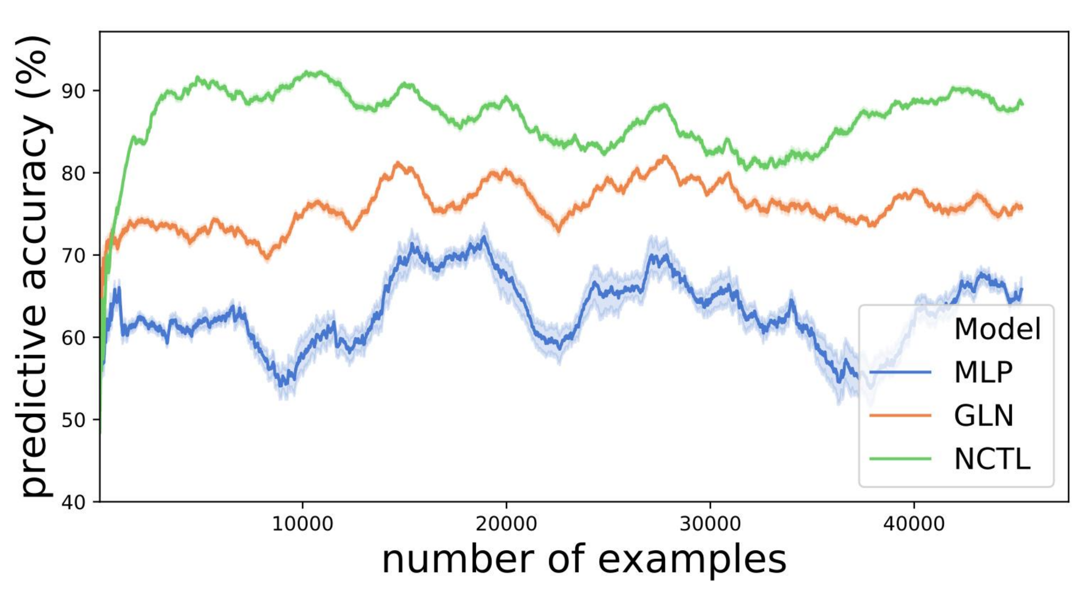
NCTL shows desirable continual learning properties: positive forward/backward transfer, easy interpretability and competitive performance against oracles.

NCTL achieves competitive or SOTA performance across

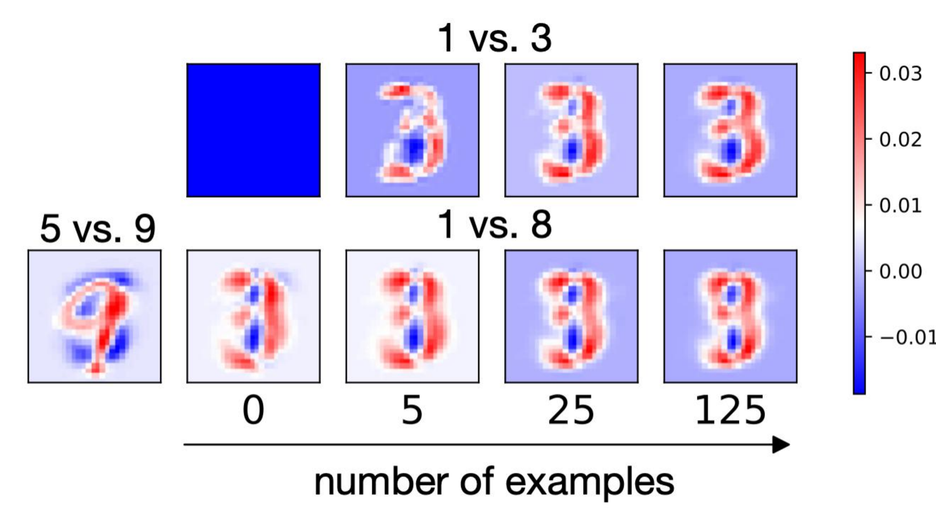
1. Split Fashion MNIST, Spit/Permuted MNIST solving a more difficult variant of the problem without access to task changepoints and task boundaries.
2. Real-world dataset *Electricity*.

Method	Experience Replay	Task Boundaries	Split MNIST	Permuted MNIST	Reference
NCTL			**95.07** $\pm$ 0.02	**95.27** $\pm$ 0.00	[ours]
EWC		✓	58.85 $\pm$ 2.59	91.04 $\pm$ 0.48	[KPR+17]
Online EWC		✓	57.33 $\pm$ 1.44	92.51 $\pm$ 0.39	[Hus18]
SI		✓	64.76 $\pm$ 3.09	93.94 $\pm$ 0.45	[ZPG17]
MAS		✓	68.57$\pm$ 6.85	94.08 $\pm$ 0.43	[ABE+18]
LwF		✓	**71.02** $\pm$ 1.26	**72.64** $\pm$ 0.52	[LH17]
GEM	✓	✓	96.16 $\pm$ 0.35	96.19 $\pm$ 0.11	[LPR17]
DGR	✓	✓	95.74 $\pm$ 0.23	95.09 $\pm$ 0.04	[SLKK17]
RtF	✓	✓	**97.31** $\pm$ 0.11	**97.06** $\pm$ 0.02	[vdVT18]
Offline (upper bound)			98.59 $\pm$ 0.15	97.90 $\pm$ 0.09	[HLRK18]

Split/Permuted MNIST



Real-world dataset *Electricity* (Elec2-3)



Easy Interpretability via Linear Saliency

Join the GLN revolution!

[1] Veness, Joel, et al. "Online learning with gated linear networks." arXiv preprint arXiv:1712.01897 (2017)

[2] Veness, Joel, et al. "Gated linear networks." arXiv preprint arXiv:1910.01526 (2019)

[3] Milan, Kieran, et al. "The forget-me-not process." Advances in Neural Information Processing Systems (2016)

[4] **Poster 17872** Budden, David, et al. "Gaussian Gated Linea Networks." arXiv preprint arXiv: 2006.0596

[5] **Poster 18607** Sezener, Eren, et al "Online Learning in Contextual Bandits using Gated Linear Networks." arXiv preprint arXiv:2002.11611 (2020)