Topic Modeling as Multi-Objective Contrastive Optimization

Thong Nguyen, Xiaobao Wu, Xinshuai Dong,
Cong-Duy Nguyen, See-Kiong Ng, Luu Anh Tuan

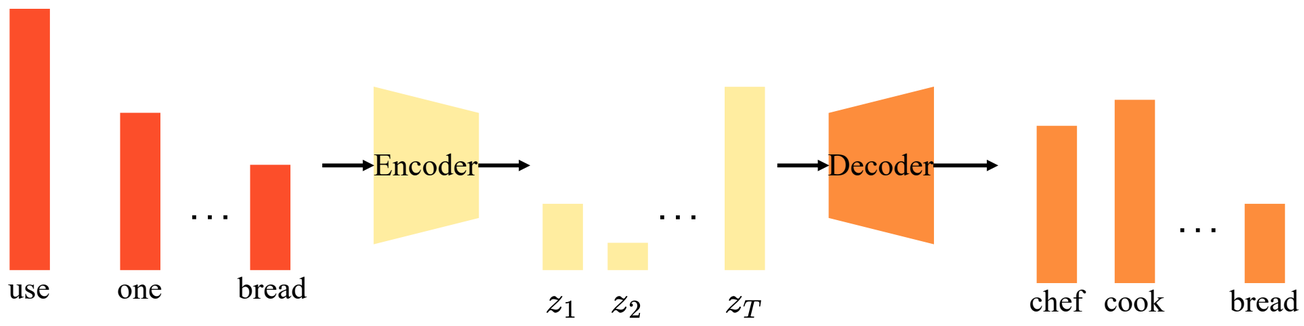
ICLR International Conference On Learning Representations

Paper Code

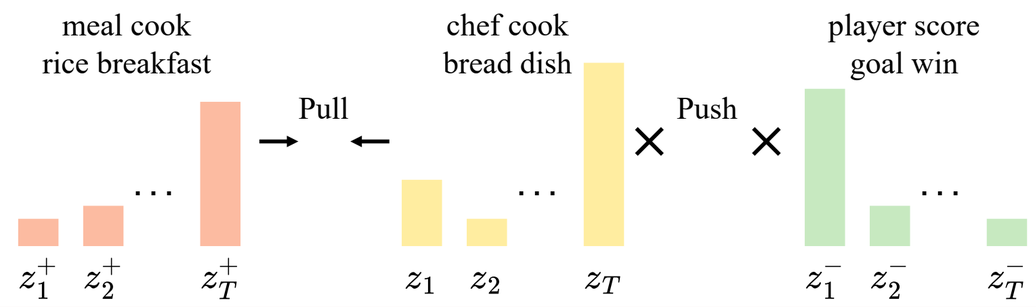
National University of Singapore
Nanyang Technological University Singapore
Carnegie Mellon University Pittsburgh Pennsylvania 1900

Contrastive Learning for Topic Modeling

Topic Modelling (TM)

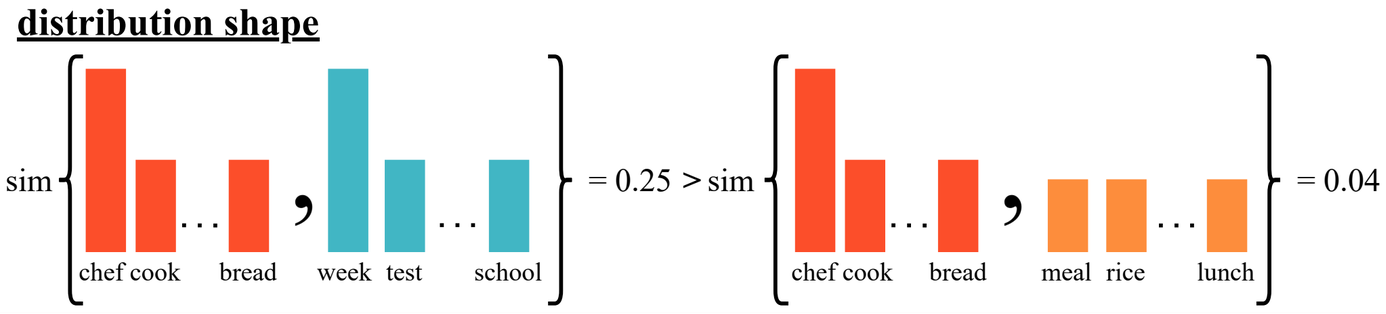


Instance-based Contastive Learning (Instance-based CL)



Problem: Instance-based Contrastive Learning Overwhelms Topic Modeling

1) Instance CL may make topic model focus on low-level feature, *e.g.* distribution shape

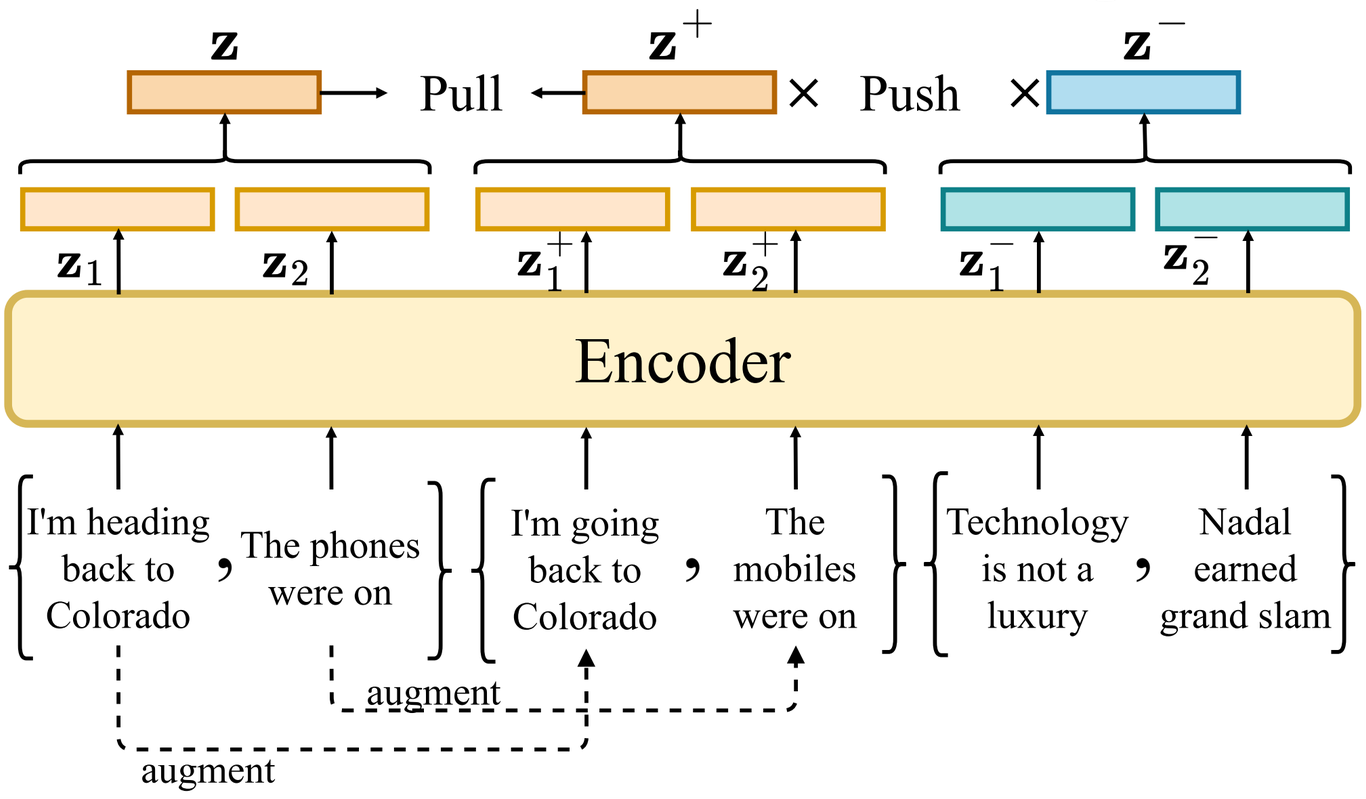


2) Instance CL loss may make topic model focus on peculiar words.

sim (shuttle land on planet, job career ask development) = 0.0093

sim (shuttle land on planet **zeppelin**, job career ask development **zeppelin**) = 0.9741

Set-based Contrastive Learning



Parameter α to balance CL and TM

Optimization objective: $\alpha \cdot L_{\text{CL}} + (1 - \alpha) \cdot L_{\text{TM}}$

We find α that achieves a balance through solving this optimization problem:

$$\min_{\alpha} \left\{ ||\alpha \nabla_\theta L_{\text{CL}}(\theta) + (1 - \alpha) \nabla_\theta L_{\text{TM}}(\theta, \phi)||_2^2, \alpha \geq 0 \right\}$$

→ Solution: use derivative!

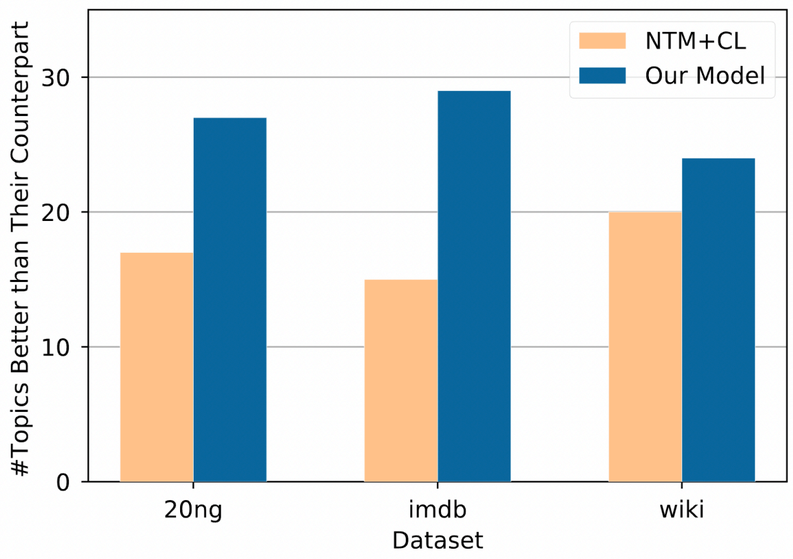
Then, we optimize the objective which is weighted by the found α.

Experiments

Quantitative Results

Method	20NG				IMDb			
	$T = 50$		$T = 200$		$T = 50$		$T = 200$	
	NPMI	TD	NPMI	TD	NPMI	TD	NPMI	TD
NTM	0.283±0.004	0.734±0.009	0.277±0.003	0.686±0.004	0.170±0.008	0.777±0.021	0.169±0.003	0.690±0.015
ETM	0.305±0.006	0.776±0.022	0.264±0.002	0.623±0.002	0.174±0.001	0.805±0.019	0.168±0.001	0.687±0.007
DVAE	0.320±0.005	0.824±0.017	0.269±0.003	0.786±0.005	0.183±0.004	0.836±0.010	0.173±0.006	0.739±0.005
BATM	0.314±0.003	0.786±0.014	0.245±0.001	0.623±0.008	0.065±0.008	0.619±0.016	0.090±0.004	0.652±0.008
W-LDA	0.279±0.003	0.719±0.026	0.188±0.001	0.614±0.002	0.136±0.007	0.692±0.016	0.095±0.003	0.666±0.009
SCHOLAR	0.319±0.007	0.788±0.008	0.263±0.002	0.634±0.006	0.168±0.002	0.702±0.014	0.140±0.001	0.675±0.005
SCHOLAR + BAT	0.324±0.006	0.824±0.011	0.272±0.002	0.648±0.009	0.182±0.002	0.825±0.008	0.175±0.003	0.761±0.010
NTM+CL	0.332±0.006	0.853±0.005	0.277±0.003	0.699±0.004	0.191±0.004	0.857±0.010	0.186±0.002	0.843±0.008
HyperMiner	0.305±0.006	0.613±0.023	0.254±0.002	0.646±0.004	0.182±0.004	0.485±0.009	0.177±0.002	0.658±0.012
WeTe	0.304±0.005	0.749±0.018	0.254±0.001	0.742±0.005	0.167±0.004	0.831±0.010	0.163±0.005	0.738±0.008
TSCTM	0.271±0.007	0.668±0.019	0.226±0.001	0.662±0.006	0.149±0.003	0.741±0.008	0.145±0.002	0.658±0.012
Our model	**0.340**±0.005	**0.913**±0.019	**0.291**±0.003	**0.905**±0.004	**0.200**±0.007	**0.916**±0.008	**0.197**±0.003	**0.892**±0.007

We have more better topics than baseline model.



Qualitative Results

Dataset	Method	NPMI	Topic
20NG	NTM+CL	0.2766	mouse monitor orange gateway video apple screen card port vga
	Our Model	0.3537	vga monitor monitors colors video screen card mhz cards color
IMDb	NTM+CL	0.1901	seagal ninja martial arts zombie zombies jet fighter flight helicopter
	Our Model	0.3143	martial arts seagal jackie chan kung hong ninja stunts kong
Wiki	NTM+CL	0.1070	architectural castle architect buildings grade historic coaster roller sculpture tower
	Our Model	0.2513	century building built church house site castle buildings historic listed