Theoretical background

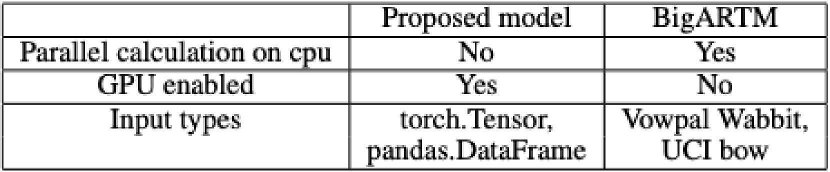
The proposed problem is the problem of stochastic matrix factorization. For it to have a unique solution, it has to be regularized as follows:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta} \tag{1}$$

Where $R(\Phi, \Theta)$ — additive regularization term. This is the classic problem of ARTM. BigARTM authors proposed a solution using expectation-maximization algorithm:

$$\begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{norm} \\ \phi_{wt} = \underset{w \in W}{norm} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{norm} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} \end{cases} \tag{2}$$

Quick comparison to sota-model



	Proposed model	BigARTM
Parallel calculation on cpu	No	Yes
GPU enabled	Yes	No
Input types	torch.Tensor, pandas.DataFrame	Vowpal Wabbit, UCI bow

Table 1: Brief comparison of BigARTM and proposed model

Model architecture

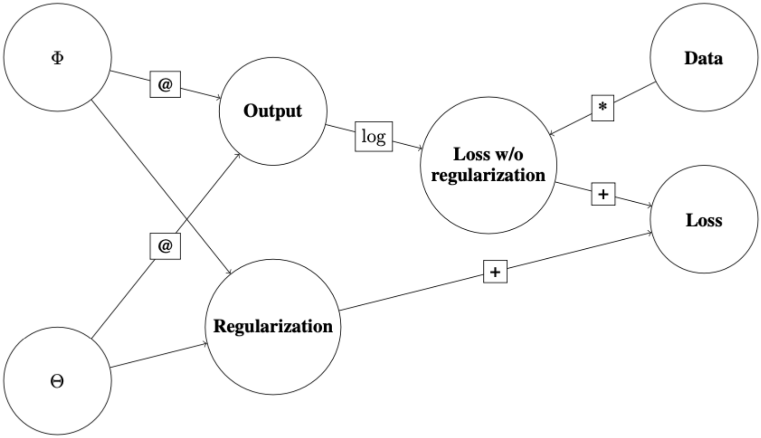
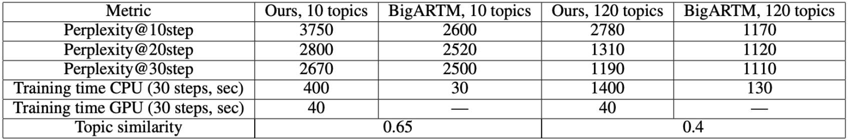


Figure 1: Calculation graph

Experiments

We evaluate the proposed model on the 20newsgroups dataset, which contains a collection of approximately 20,000 newsgroup documents, partitioned nearly evenly across 20 different newsgroups. We extracted approximately 70,000 unique words as vocabulary in bag-of-words format.



Metric	Ours, 10 topics	BigARTM, 10 topics	Ours, 120 topics	BigARTM, 120 topics
Perplexity@10step	3750	2600	2780	1170
Perplexity@20step	2800	2520	1310	1120
Perplexity@30step	2670	2500	1190	1110
Training time CPU (30 steps, sec)	400	30	1400	130
Training time GPU (30 steps, sec)	40	—	40	—
Topic similarity	0.65		0.4	

Рис.: Quantiative results

Qualitative Experiments

We analyze our outcomes in terms of top-k words, representing the topic, focusing on models with 10 topics. These particular examples demonstrate the same logic of clusterization in both models.

topic 1: max, q, r, g, p
topic 2: one, would, say, people, write
topic 3: game, team, line, subject, organization
topic 4: would, people, write, gun, article
topic 5: god, hell, atheist, line, subject
topic 6: x, file, use, window, program
topic 7: say, armenian, people, one, go
topic 8: line, subject, get, organization, car
topic 9: space, organization, subject, line, db
topic 10: line, subject, organization, use, university

topic 1: max, q, r, g, p
topic 2: line, one, get, subject, use
topic 3: god, would, say, one, people
topic 4: line, subject, organization, university, article
topic 5: game, team, line, drive, subject
topic 6: say, armenian, one, go, people
topic 7: x, file, line, use, subject
topic 8: use, line, subject, organization, window
topic 9: would, people, write, get, article
topic 10: use, key, system, data, space