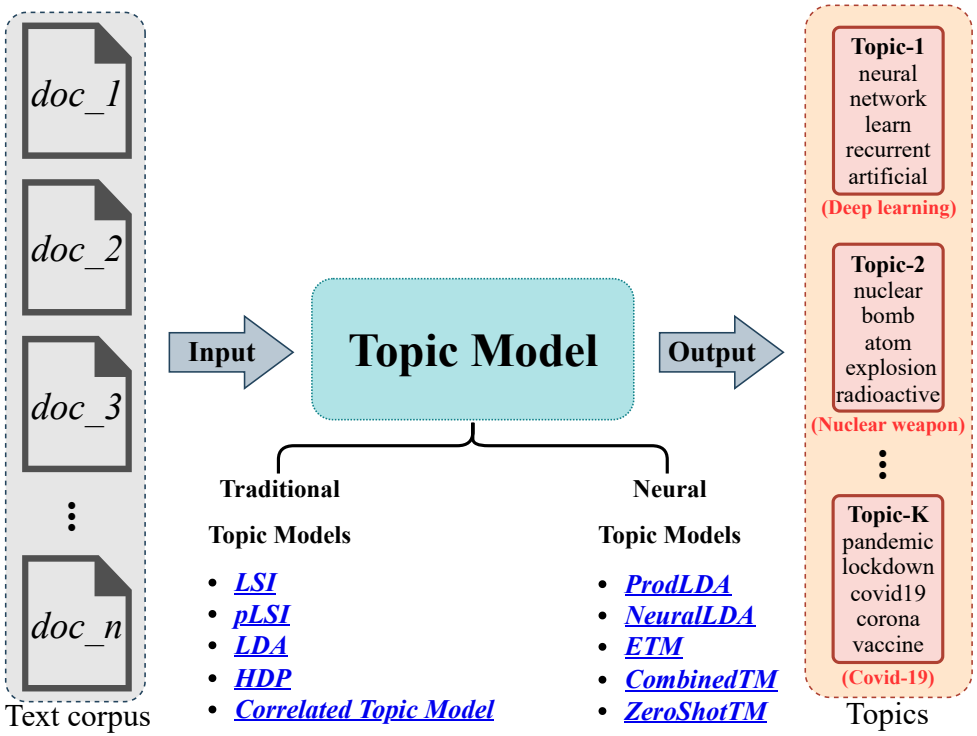
Improving Neural Topic Models with Wasserstein Knowledge Distillation

Suman Adhya and Debarshi Kumar Sanyal

Indian Association for the Cultivation of Science
Jadavpur, Kolkata 700032, INDIA

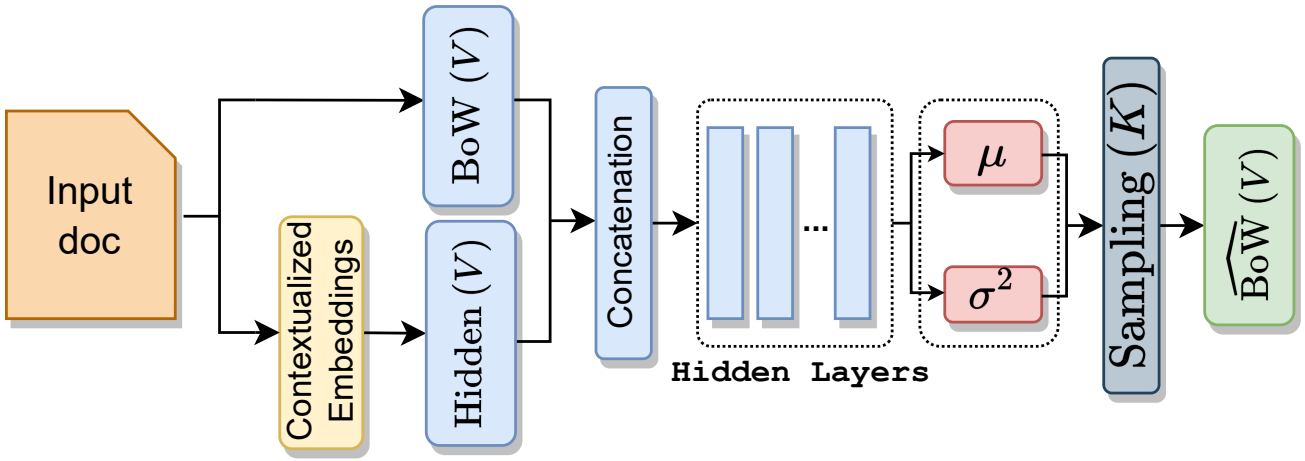
ECIR 2023

What is a Topic Model?



- **Type**: Unsupervised learning;
- **Input**: Set of documents;
  Output: Set of topics;
- **Topic**: Distribution over the words;
- **Use cases**:
  1. Document clustering;
  2. Text classification;
  3. Information retrieval;

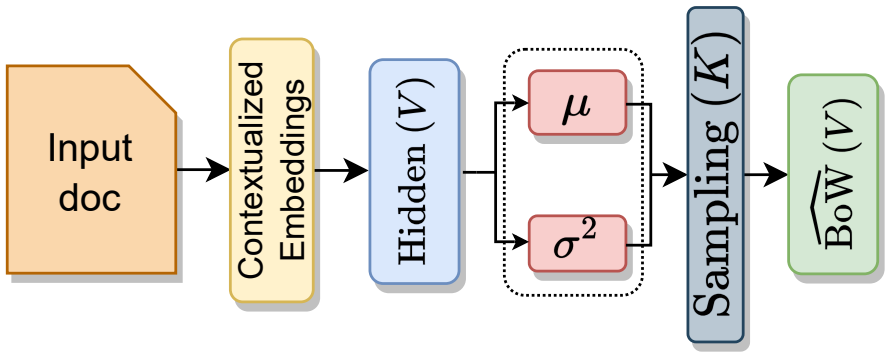
CombinedTM



CombinedTM [1].

- **Benefit**: **Performance** ↑
- **Drawback**: ↑ **Network complexity** ⇒ ↑ **Resource requirements**

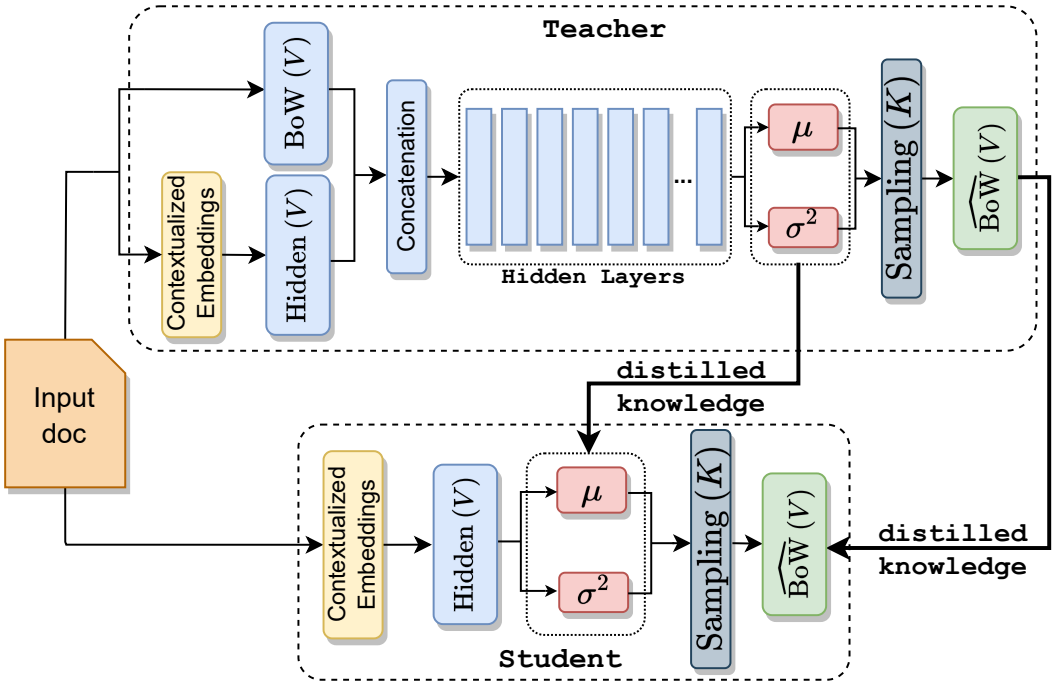
ZeroShotTM



ZeroShotTM [2] with one hidden layer.

- **Benefit**: **Model size** ↓
- **Drawback**: **Performance** ↓

Proposed Distillation Framework



KD framework.

- **Teacher** (T): **CombinedTM**;
- **Student** (S'): **ZeroShotTM**;
- The **pre-trained model** T will be used to train the model S' with knowledge distillation.

Student's Loss Function

The total loss for the student (S') is defined as:

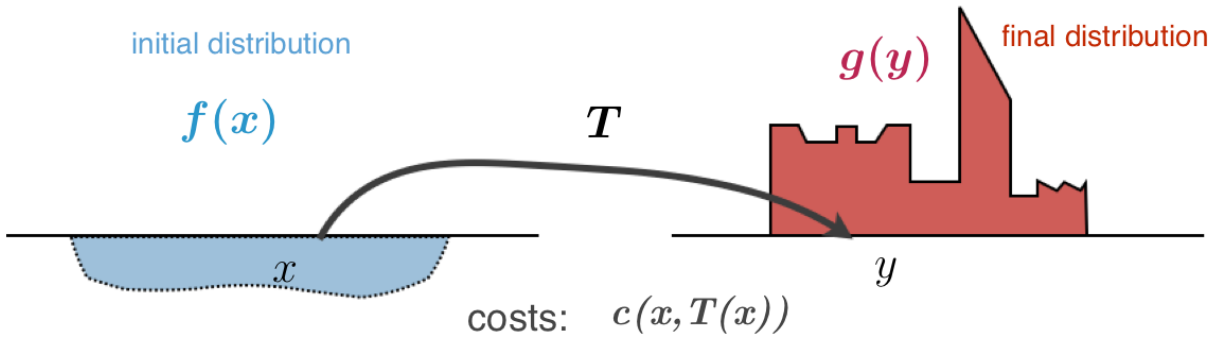
$$\mathcal{L}_{S'} = (1-\alpha)\mathcal{L}_{\text{VAE}} + \alpha\mathcal{L}_{\text{KD}}$$

Where:

- $\mathcal{L}_{\text{VAE}}$: **VAE loss** function.
- $\mathcal{L}_{\text{KD}}$: **KD loss** function.
- $\alpha \in [0, 1]$: **hyperparameter**.

Wasserstein Metric

Goal: Distance function defined between two probability distributions.



Interpretation: The minimum cost of moving a pile of dirt in the shape of one probability distribution to the shape of another distribution.

KD Loss Function

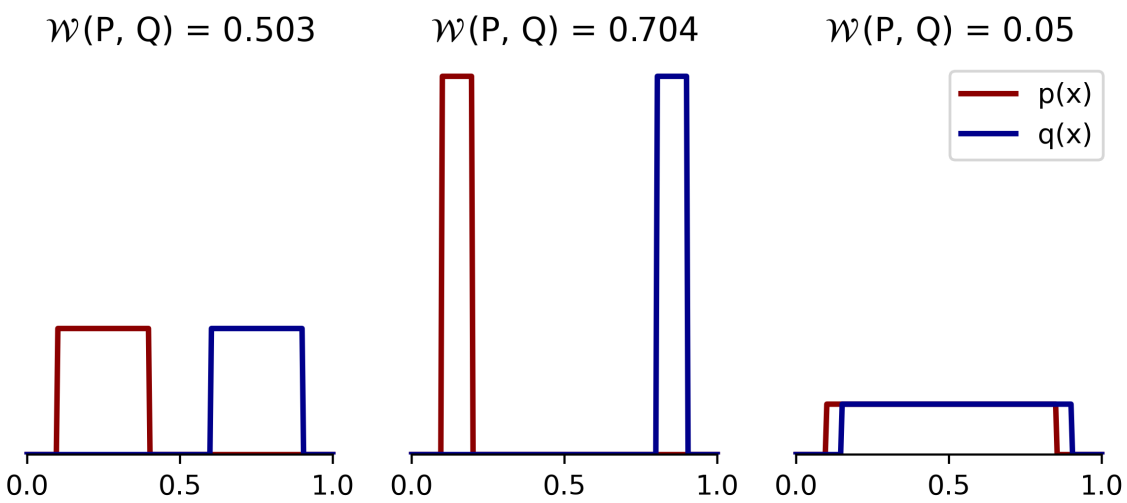
The $\mathcal{L}_{\text{KD}}$ has two components:

$$\mathcal{L}_{KD} = \mathcal{L}_{\text{KD-2W}} + t^2 \mathcal{L}_{\text{KD-CE}}$$

- $\mathcal{L}_{\text{KD-2W}}$: Squared of the 2-Wasserstein distance between the latent distributions learned by the two models T and S'.
- $\mathcal{L}_{\text{KD-CE}}$: Cross-entropy between the soft labels produced by T and S'.
- $t \in \{1, 2, \ldots, 5\}$: Softmax temperature (*hyperparameter*).

Why Choose Wasserstein over KL?

KL divergence is sensitive to small differences in the distributions.



Source: Alex Williams. *"A Short Introduction to Optimal Transport and Wasserstein Distance"* 09 Oct 2020.

Unlike KL divergence which is **infinity in all three cases**, the Wasserstein distances in these examples are **finite and intuitive**.

Experimental Setup

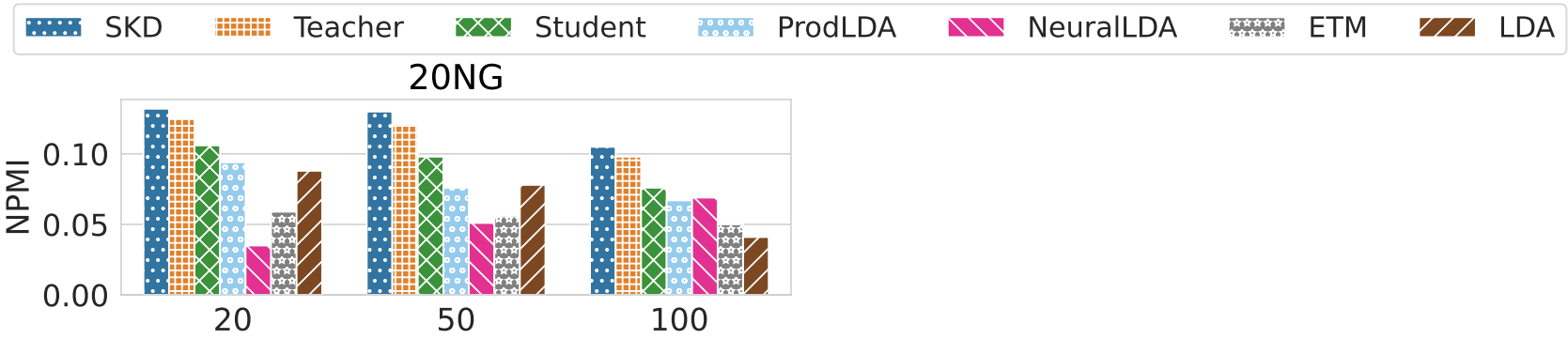
All experiments are performed using **OCTIS** [6].

- **Datasets**:

Name	Type	#Docs
20NG	Newsgroups posts on 20 topics	16,309
M10	Scientific publications	8,355

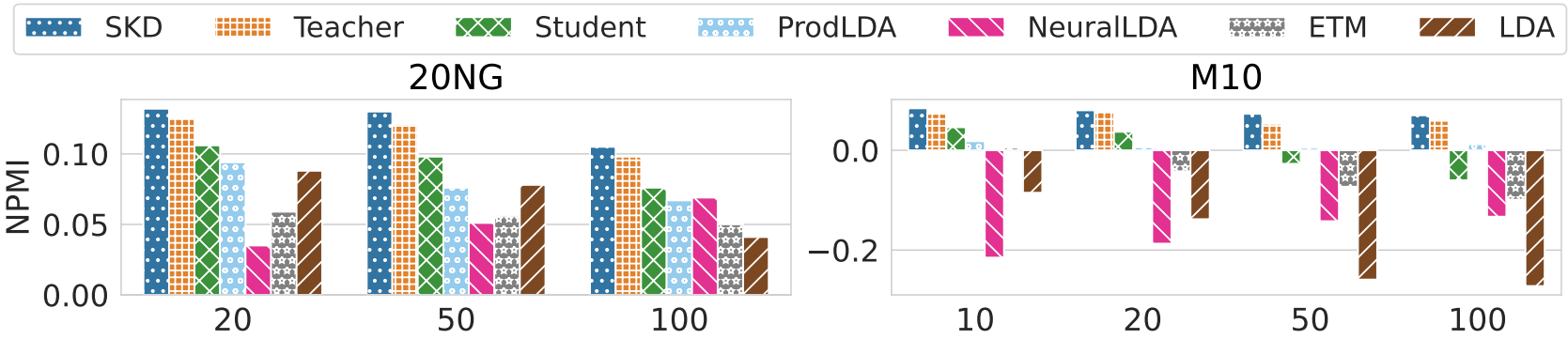
- **Baselines**:
  - **CombinedTM** [1], **ZeroShotTM** [2], **ProdLDA** [5], **NeuralLDA** [5], **ETM** [4], **LDA** [3]
- **Evaluation metrics**:
  - Topic coherence (topic-words relevancy): **NPMI**, **CV**.

Quantitative Evaluation



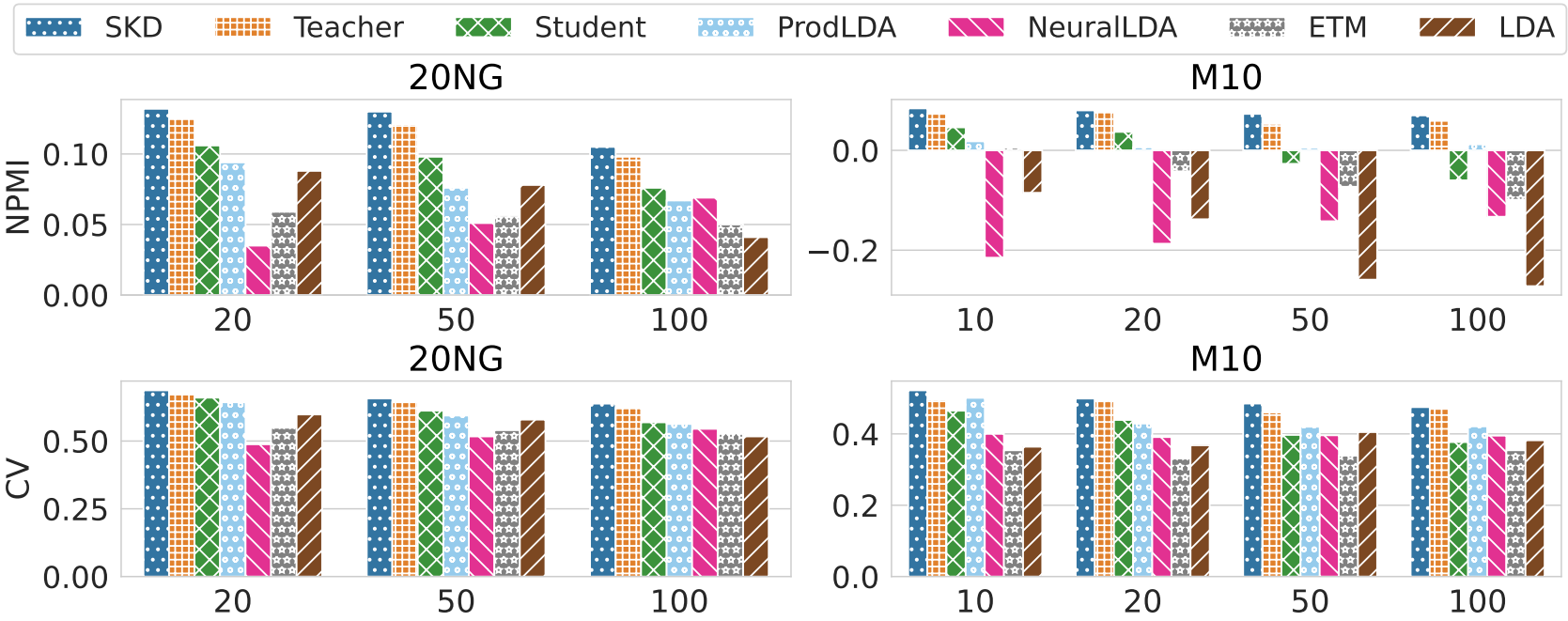
Coherence scores (**NPMI** and **CV**) for different topic models on two datasets: **20NG** and **M10**. The X-axis is marked with the topic counts used for each dataset.

Quantitative Evaluation



Coherence scores (**NPMI** and **CV**) for different topic models on two datasets: **20NG** and **M10**. The X-axis is marked with the topic counts used for each dataset.

Quantitative Evaluation



Coherence scores (**NPMI** and **CV**) for different topic models on two datasets: **20NG** and **M10**. The X-axis is marked with the topic counts used for each dataset.

Qualitative Evaluation

Model	ID	Topics
T	0	gun, law, firearm, crime, weapon, assault, amendment, state, police, permit
	11	russian, turkish, people, village, genocide, armenian, muslim, population, greek, army
	17	oil, engine, ride, front, road, chain, bike, motorcycle, water, gas
S	0	law, people, state, government, gun, amendment, constitution, firearm, crime, privacy
	1	armenian, village, soldier, soviet, muslim, troop, turkish, russian, genocide, land
	17	engine, car, mile, ride, bike, oil, front, wheel, motorcycle, tire
SKD	0	gun, law, weapon, firearm, amendment, crime, bill, assault, constitution, police
	11	turkish, genocide, armenian, russian, village, population, israeli, war, attack, muslim
	17	ride, engine, car, bike, motorcycle, front, oil, motor, road, seat

- Some aligned topics of **T**, **S**, and **SKD** from **20NG** for **20 topics**.
- The words in a topic from **S** or **SKD** that are shared with the corresponding topic in **T** are highlighted.
- **SKD** displays more word overlap than **S** with the topics of **T**.

Model Compression

The sizes of all the models depend on:

1. Dimension of contextualized embeddings.
2. Number and size of hidden layers.
3. Number of topics.
4. Vocabulary size.

$$|\mathbf{S}| = |\mathbf{SKD}| < |\mathbf{T}|$$

For **20 topics** in **20NG**, the reduction in model size is 55.4%.
In general, the compression ranged from 37.6% to 56.3%.

Conclusion & Future Directions

Conclusion: Proposed a 2-Wasserstein loss-based knowledge distillation framework to compress the neural topic models without compromising the performance.

Future work:

- Study it analytically.
- Apply it to distill knowledge across other neural topic models.

References

[1] Federico Bianchi, Silvia Terragni, and Dirk Hovy. "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *ACL-IJCNLP* (2021).

[2] Federico Bianchi et al. "Cross-lingual contextualized topic models with zero-shot learning". In: *EACL* (2021).

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet allocation". In: *JMLR* (2003).

[4] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *TACL* (2020).

[5] Akash Srivastava and Charles Sutton. "Autoencoding Variational Inference For Topic Models". In: *ICLR* (2017).

[6] Silvia Terragni et al. "OCTIS: Comparing and Optimizing Topic models is Simple!" In: *EACL: System Demonstrations* (2021).

Thank you all for listening...

Code: `https://github.com/AdhyaSuman/CTMKD`