Department of Mathematical Science, KAIST

Neural Topic Modeling with Continual Lifelong Learning

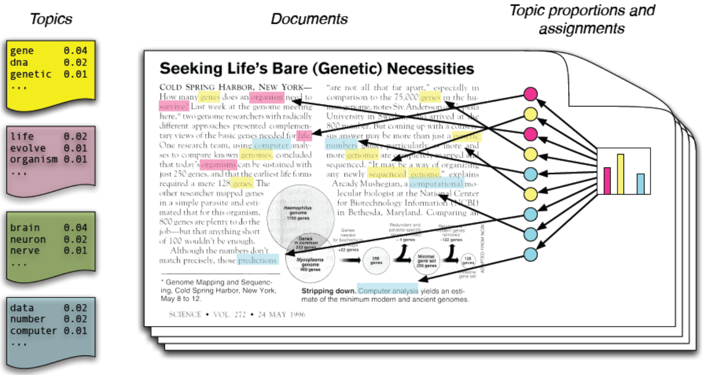
Jinhwan Suk

October 27, 2020

Topic Modeling

Introduction : LDA, 2003

- Probabilistic topic models are used to extract topics from text collections.
- Information retrieval, document classification, or summarization.



Topic Modeling

Introduction : LDA, 2003

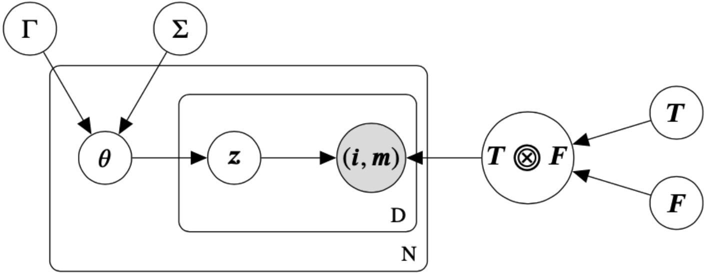


Figure 1: Graphical model for structural topic modelling

Restricted Boltzmann Machines (RBMs)

Binary RBMs

The most common form of RBM has binary hidden nodes and binary visible nodes. The joint distribution has following form:

$$p(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}))$$

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v} - \boldsymbol{b}^T \boldsymbol{v} - \boldsymbol{c}^T \boldsymbol{h} \triangleq -\boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}$$

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{v}, \boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}))$$

$\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c})$ are all the parameters. It is common to use **stochastic gradient descent**, since RBMs often have many parameters.

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax} p(\boldsymbol{v}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}))$$

Restricted Boltzmann Machines (RBMs)

Binary RBMs : Learning

$$F(\boldsymbol{v}) := -\log \tilde{p}(\boldsymbol{v})$$

$$\tilde{p}(\boldsymbol{v}) = \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})) = \sum_{\boldsymbol{h}} \exp\left(\sum_{i=1}^{D}\sum_{j=1}^{K} W_{ij} v_i h_j\right)$$

$$= \sum_{\boldsymbol{h}} \prod_{j=1}^{K} \exp\left(\sum_{i=1}^{D} W_{ij} v_i h_j\right) = \prod_{j=1}^{K} \sum_{h_j \in \{0,1\}} \exp\left(\sum_{i=1}^{D} W_{ij} v_i h_j\right)$$

$$= \prod_{j=1}^{K} \left(1 + \exp\left(\sum_{i=1}^{D} W_{ij} v_i\right)\right)$$

$$\ell(\boldsymbol{\theta}) = \frac{1}{D}\sum_{i=1}^{D} \log p(v_i;\, \boldsymbol{\theta}) = -\frac{1}{D}\sum_{i=1}^{D} F(v_i;\, \boldsymbol{\theta}) - \log Z(\boldsymbol{\theta})$$

Since

$$\frac{\partial}{\partial w_{ij}} F(\boldsymbol{v}) = -\mathbb{E}\left[v_i h_j \mid \boldsymbol{v}, \boldsymbol{\theta}\right]$$

Hence

$$\frac{\partial}{\partial w_{ij}} \ell(\boldsymbol{\theta}) = \frac{1}{D} \sum_{i=1}^{D} \mathbb{E}\left[v_i h_j \mid \boldsymbol{v}, \boldsymbol{\theta}\right] - \mathbb{E}\left[v_i h_j \mid \boldsymbol{\theta}\right] \tag{1}$$

The conditional expectation $\mathbb{E}[\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}]$ is sigm($\boldsymbol{W}^T\boldsymbol{v}$). Approximating $\mathbb{E}\left[v_i h_j \mid \boldsymbol{\theta}\right]$ ..?

$\rightarrow$ Gibbs Sampling, Contrastive divergence (CD), Persistent CD

$\rightarrow$ The complexity of RSM is $\mathcal{O}(V)$

Replicated Softmax (RSM) : an Undirected Topic Model

Salakhutdinov and Hinton, 2009

- **RSM** is the extension of the binary RBM to categorical variable.

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{D}\sum_{j=1}^{H}\sum_{k=1}^{V} W_{ij}^{k} v_i^k h_j - \sum_{i=1}^{D}\sum_{k=1}^{V} v_i^k b_i^k - \sum_{j=1}^{H} h_j a_j \tag{2}$$

- The conditional distributions are given by

$$p(v_i^k = 1 \mid \boldsymbol{h}) = \frac{\exp(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k)}{\sum_{q=1}^{K} \exp(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q)}$$

$$p(h_j = 1 \mid \boldsymbol{v}) = \sigma\left(a_j + \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k W_{ij}^k\right)$$

Assuming we can ignore the order of the words, $\forall\, i,\ b = b_i,\ W_j^k = W_{ij}^k$

Replicated Softmax (RSM) : an Undirected Topic Model

Salakhutdinov and Hinton, 2009

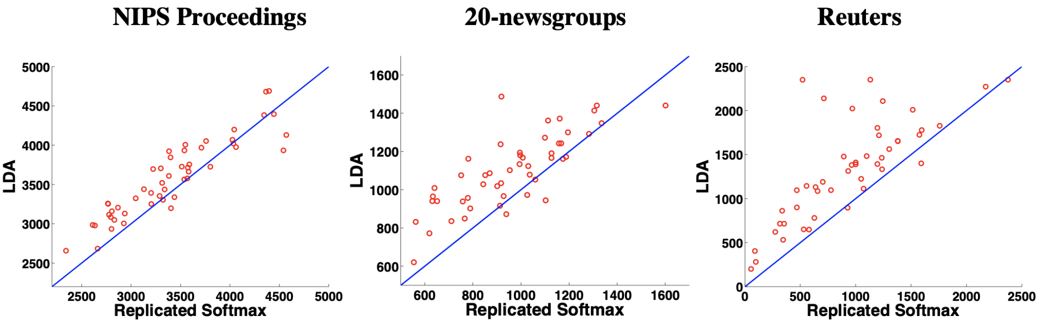
- Document length is not matter.
- Instead of representing documents as distributions over topics, relies on a **distributed representation** of the documents. (e.g.) (0.5, 0.3, 0.3)
- Evaluation metric : Perplexity score (PPL)

$$PPL = p(\boldsymbol{v})^{-\frac{1}{D}}$$

Computed using "Annealed Importance Sampling"

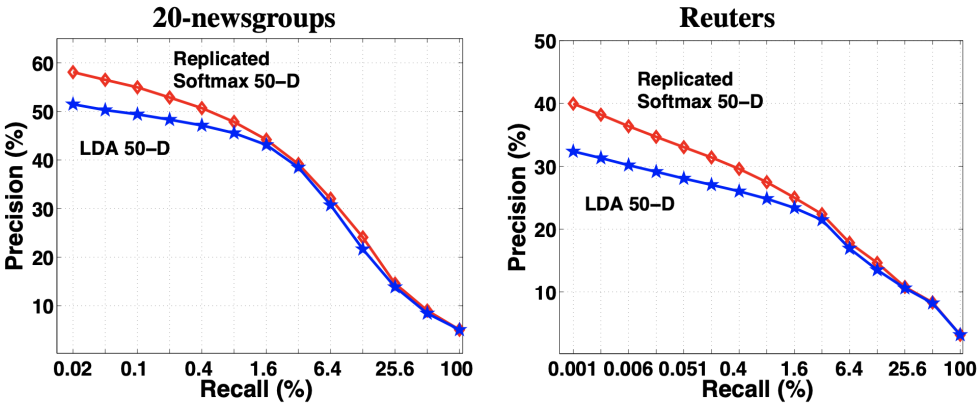
Replicated Softmax (RSM) : an Undirected Topic Model

Salakhutdinov and Hinton, 2009



Replicated Softmax (RSM) : an Undirected Topic Model

Salakhutdinov and Hinton, 2009



LDA vs RSM

Comparision

- Interpretability : LDA $>$ RSM
- Complexity : LDA $<$ RSM$(\mathcal{O}(VH + DH))$
- Predictability : LDA $<$ RSM

A Neural Autoregressive Topic Model

H. Larochelle and S. Lauly, 2012

- **NADE** is a generative model over binary observations $\{0,1\}^D$

$$p(\boldsymbol{v}) = \prod_{i=1}^{D} p(v_i \mid \boldsymbol{v}_{<i}), \quad p(v_i = 1 \mid \boldsymbol{v}_{<i}) = \sigma\left(b_i + \boldsymbol{V}_{i,:}\boldsymbol{h}_i\right), \quad \boldsymbol{h}_i = \sigma\left(\boldsymbol{c} + \boldsymbol{W}_{:,<i}\boldsymbol{v}_{<i}\right)$$

- The parameters $\{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{W}, \boldsymbol{V}\}$ are learned by minimizing NLL with SGD.
- **DocNADE** is a model over V-observations $\{1, \ldots, V\}^D$.

$$p(v_i = w|\boldsymbol{v}_{<i}) = \frac{\exp(b_w + V_{w,:}\boldsymbol{h}_i)}{\underbrace{\sum_{w'} \exp(b_w + V_{w,:}\boldsymbol{h}_i)}_{\mathcal{O}(V) \Rightarrow \text{expensive!!}}}, \quad \boldsymbol{h}_i = \sigma\left(\boldsymbol{c} + \sum_{k<i} \boldsymbol{W}_{:,v_k}\right)$$

A Neural Autoregressive Topic Model

H. Larochelle and S. Lauly, 2012

- *NADE* and *DocNADE* were directly inspired from the RBM.
- The distribution of an RBM could be written as

$$p(\boldsymbol{v}) = \prod_{i=1}^{D} p(v_i|\boldsymbol{v}_{<i}) = \prod_{i=1}^{D} \frac{p(v_i, \boldsymbol{v}_{<i})}{p(\boldsymbol{v}_{<i})} = \prod_{i=1}^{D} \frac{\sum_{\boldsymbol{v}_{>i}} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))}{\sum_{\boldsymbol{v}_{\geq i}} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))}$$

- Since the conditionals are intractable, we first find an approximation

$$q(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{v}_{<i}) \approx p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{v}_{<i})$$

with mean-field assumption :

$$q(v_i, \boldsymbol{v}_{>i}, \boldsymbol{h}|\boldsymbol{v}_{<i}) = \mu_i(i)^{v_i}(1-\mu_i(i))^{1-v_i} \prod_{j>i} \mu_j(i)^{v_j}(1-\mu_j(i))^{1-v_j} \prod_{k} \tau_k(i)^{h_k}(1-\tau_k(i))^{1-h_k}$$

A Neural Autoregressive Topic Model

H. Larochelle and S. Lauly, 2012

Minimizing $\mathcal{D}_{KL}(q(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{v}_{<i})||p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{v}_{<i}))$, we have

$$\tau_k(i) = \sigma\left(c_k + \sum_{j \geq i} W_{kj}\mu_j(i) + \sum_{j<i} W_{kj}v_j\right)$$

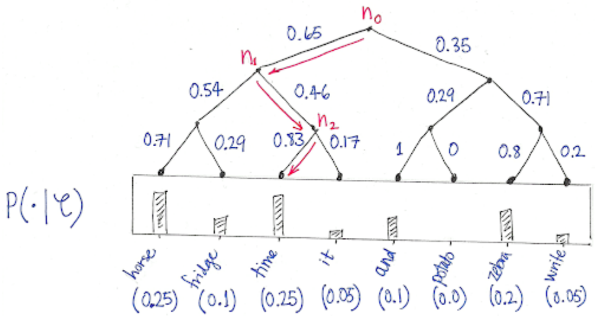
$$\mu_j(i) = \sigma\left(b_j + \sum_{k} W_{kj}\tau_k(i)\right), \quad \forall j \geq i$$

With initial $\mu_j(i)$, $j \geq i$ to be 0, iterate only once. We can rewrite as follows :

$$p(v_i = 1|\boldsymbol{v}_{<i}) = \sigma\left(b_i + \left(W^\top\right)_{i,:} h_i\right), \quad h_i = \sigma\left(c + W_{\cdot,<i}\boldsymbol{v}_{<i}\right)$$

A Neural Autoregressive Topic Model

H. Larochelle and S. Lauly, 2012



$$p(v_i = w|\boldsymbol{v}_{<i}) = \underbrace{\prod_{m=1}^{|\boldsymbol{\pi}(v_i)|} p\left(\pi(v_i)_m|\boldsymbol{v}_{<i}\right)}_{\mathcal{O}(\log V)}, \quad p\left(\pi(v_i)_m = 1|\boldsymbol{v}_{<i}\right) = \sigma\left(b_{l(v_i)_m} + \boldsymbol{V}_{l(v_i)_m,:}\boldsymbol{h}_i\right)$$

A Neural Autoregressive Topic Model

H. Larochelle and S. Lauly, 2012

Generative Model Evaluation

Data Set	LDA (50)	LDA (200)	Replicated Softmax (50)	DocNADE (50)	DocNADE St. Dev
20 Newsgroups	1091	1058	953	**896**	6.9
RCV1-v2	1437	1142	988	**742**	4.5

- perplexity per word score :

$$PPW = \exp\left(-\frac{1}{N}\sum_{t}\frac{1}{|D^t|}\log p(D^t)\right)$$

Document Retrieval Evaluation

Hidden unit topics			
jesus	shuttle	season	encryption
atheism	orbit	players	escrow
christianity	lunar	nhl	pgp
christ	spacecraft	league	crypto
athos	nasa	braves	nsa
atheists	space	playoffs	rutgers
bible	launch	rangers	clipper
christians	saturn	hockey	secure
sin	billion	pitching	encrypted
atheist	satellite	team	keys

Table 2: The five nearest neighbors in the word representation space learned by DocNADE.

weapons	**medical**	**companies**	**define**	**israel**	**book**	**windows**
weapon	treatment	demand	defined	israeli	reading	dos
shooting	medecine	commercial	definition	israelis	read	microsoft
firearms	patients	agency	refer	arab	books	version
assault	process	company	make	palestinian	relevent	ms
armed	studies	credit	examples	arabs	collection	pc

Document Informed Neural Autoregressive Topic Models with Distributional Prior

P. Gupta et al., AAAI, 2012

- In DocNADE, to predict the word v_i, each hidden layer h_i takes $\boldsymbol{v}_{<i}$ as the input.
- It doesn't take into account the following words $\boldsymbol{v}_{>i}$
- They extended DocNADE to incorporate full contextual information

Document Informed Neural Autoregressive Topic Models with Distributional Prior

P. Gupta et al., AAAI, 2012

- In DocNADE, to predict the word v_i, each hidden layer h_i takes $\boldsymbol{v}_{<i}$ as the input.
- It doesn't take into account the following words $\boldsymbol{v}_{>i}$
- They extended DocNADE to incorporate full contextual information
- Only powerful for long texts and corpora with many documents.
- They incorporated *pre-trained word embeddings*, E.

Document Informed Neural Autoregressive Topic Models with Distributional Prior

P. Gupta et al., AAAI, 2012

$$h_i^{\ell \to r} = \sigma\left(c^{\ell \to r} + \sum_{k<i} W_{:,v_k} + \lambda \sum_{k<i} E_{:,v_k}\right)$$

$$h_i^{r \to \ell} = \sigma\left(c^{r \to \ell} + \sum_{k>i} W_{:,v_k} + \lambda \sum_{k>i} E_{:,v_k}\right)$$

$$\log p(\boldsymbol{v}) = \frac{1}{2} \sum_{i=1}^{D} [\log p(v_i | \boldsymbol{v}_{<i}) + \log p(v_i | \boldsymbol{v}_{>i})]$$

Neural Topic Modeling with Continual Lifelong Learning

P. Gupta et al., ICML 20'

- A stream of document collection $\boldsymbol{S} = \{\Omega^1, \Omega^2, \ldots, \Omega^T, \Omega^{T+1}\}$
- Mining and retaining **prior knowledge** from streams of document collections.
- Three main challenges in continual topic modeling :
  1. Mining prior knowledge relevant for the future task $T+1$
  2. Learning with prior knowledge
  3. Minimizing catastrophic forgetting
- All prior knowledge in this article means the embedding matrix $W \in \mathbb{R}^{K \times H}$.
  1. $W_{j,:}$ encodes j-th topic, i.e., topic-embedding(Z).
  2. $W_{:,v_i}$ corresponds to embedding of the word v_i, i.e., word-embedding(E).

Neural Topic Modeling with Continual Lifelong Learning

P. Gupta et al., ICML 20'

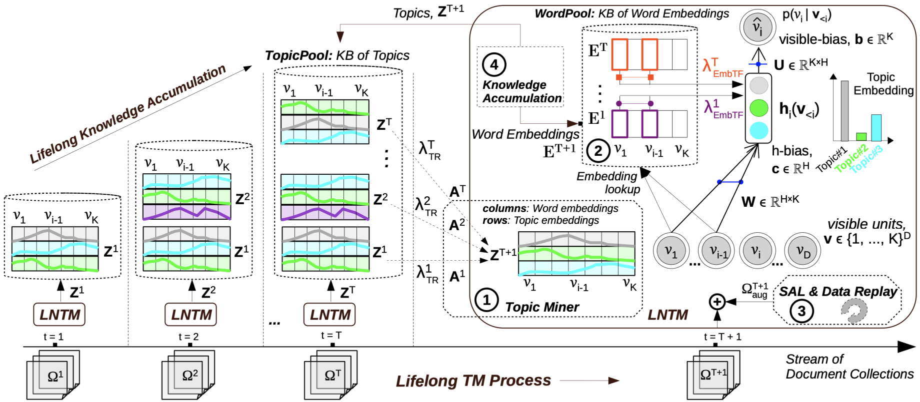


Figure 2. An illustration of the proposed Lifelong Neural Topic Modeling (LNTM) framework over s stream of document collections

- Topic Regularization

$$\mathcal{L}(\Omega^{T+1};\ \Theta^{T+1}) = \sum_{v \in \Omega^{T+1}} \mathcal{L}(v; \Theta^{T+1}) + \Delta_{TR}$$

$$\Delta_{TR} = \sum_{t=1}^{T} \lambda_{TR}^{t} \left(\underbrace{||Z^t - A^t Z^{T+1}||_2^2}_{\text{topic imitation}} + \underbrace{||U^t - P^t U||_2^2}_{\text{decoder proximity}} \right)$$

It enables jointly mining, transferring and retaining prior topics.

- Inspired by Gupta et al., 2019, we introduce prior knowledge in form of pre-trained word embeddings,

$$h_i = \sigma\left(c + \sum_{k<i} W_{:,v_k} + \sum_{k<i}\sum_{t=1}^{T} \lambda^t_{Emb} E^t_{:,v_k}\right)$$

Neural Topic Modeling with Continual Lifelong Learning

P. Gupta et al., ICML 20'

function compute-NLL ($\mathbf{v}, \boldsymbol{\Theta}$, `LNTM` $= \{\}$)
    Initialize $\mathbf{a} \leftarrow \mathbf{c}$ and $p(\mathbf{v}) \leftarrow 1$
    for word $i \in [1, ..., N]$ **do**
        $\mathbf{h}_i(\mathbf{v}_{<i}) \leftarrow g(\mathbf{a})$, where $g = \{\text{sigmoid}, \tanh\}$
        $p(v_i = w | \mathbf{v}_{<i}) \leftarrow \frac{\exp(b_w + \mathbf{U}_{w,:}\mathbf{h}_i(\mathbf{v}_{<i}))}{\sum_{w'} \exp(b_{w'} + \mathbf{U}_{w',:}\mathbf{h}_i(\mathbf{v}_{<i}))}$
        $p(\mathbf{v}) \leftarrow p(\mathbf{v}) p(v_i | \mathbf{v}_{<i})$
        Compute pre-activation at i^{th} step: $\mathbf{a} \leftarrow \mathbf{a} + \mathbf{W}_{:,v_i}$
        if `EmbTF` in `LNTM` **then**
            Get word-embedding vectors for v_i from `WordPool`:
            $\mathbf{a} \leftarrow \mathbf{a} + \sum_{t=1}^{T} \lambda_{EmbTF}^{t} \mathbf{W}_{:,v_i}^{t}$
        end if
    end for
    return $-\log p(\mathbf{v}; \boldsymbol{\Theta})$
end function

Neural Topic Modeling with Continual Lifelong Learning

P. Gupta et al., ICML 20'

input Past learning: $\{\mathbf{\Theta}^1, ..., \mathbf{\Theta}^T\}$
input `TopicPool`: $\{\mathbf{Z}^1, ..., \mathbf{Z}^T\}$
input `WordPool`: $\{\mathbf{E}^1, ..., \mathbf{E}^T\}$
parameters $\mathbf{\Theta}^{T+1} = \{\mathbf{b}, \mathbf{c}, \mathbf{W}, \mathbf{U}, \mathbf{A}^1, ..., \mathbf{A}^T, \mathbf{P}^1, ..., \mathbf{P}^T\}$
hyper-paramaters $\mathbf{\Phi}^{T+1} = \{H, \lambda^1_{LNTM}, ..., \lambda^T_{LNTM}\}$

Neural Topic Modeling:
`LNTM` $= \{\}$
Train a topic model and get PPL on test set Ω^{T+1}_{test}:
  $\text{PPL}^{T+1}, \mathbf{\Theta}^{T+1} \leftarrow$ topic-learning$(\Omega^{T+1}, \mathbf{\Theta}^{T+1})$

Lifelong Neural Topic Modeling (LNTM) framework:
`LNTM` $= \{$`EmbTF, TR, SAL`$\}$
For a document $\mathbf{v} \in \Omega^{T+1}$:
Compute loss (negative log-likelihood):
  $\mathcal{L}(\mathbf{v}|\mathbf{\Theta}^{T+1}) \leftarrow$ compute-NLL$(\mathbf{v}, \mathbf{\Theta}^{T+1}$, `LNTM`$)$
if `TR` in `LNTM` **then**
  Jointly minimize-forgetting and learn with `TopicPool`:
  $\Delta_{TR} \leftarrow \sum_{t=1}^{T} \lambda^t_{TR} \left(||\mathbf{Z}^t - \mathbf{A}^t \mathbf{Z}^{T+1}||_2^2 + ||\mathbf{U}^t - \mathbf{P}^t \mathbf{U}||_2^2 \right)$
  $\mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) \leftarrow \mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) + \Delta_{TR}$
end if
Minimize $\mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1})$ using stochastic gradient-descent
Knowledge Accumulation:
`TopicPool` $\leftarrow$ accumulate-topics$(\mathbf{\Theta}^{T+1})$
`WordPool` $\leftarrow$ accumulate-word-embeddings$(\mathbf{\Theta}^{T+1})$

Neural Topic Modeling with Continual Lifelong Learning

P. Gupta et al., ICML 20'

- Selective-Data Augmentation Learning

  $\leftarrow$ Data augmentation approaches are inefficient.

- **Step 1** *Document Distillation*

function distill-documents ($\mathbf{\Theta}^{T+1}$, PPL^{T+1}, $[\Omega^1, ..., \Omega^T]$)
  Initialize a set of selected documents: $\Omega_{aug}^{T+1} \leftarrow \{\}$
  for task $t \in [1, ..., T]$ and document $\mathbf{v}^t \in \Omega^t$ **do**
    $\mathcal{L}(\mathbf{v}^t; \mathbf{\Theta}^{T+1}) \leftarrow$ compute-NLL($\mathbf{v}^t$, $\mathbf{\Theta}^{T+1}$, $\texttt{LNTM} = \{\}$)
    $\mathrm{PPL}(\mathbf{v}^t; \mathbf{\Theta}^{T+1}) \leftarrow \exp(\frac{\mathcal{L}(\mathbf{v}^t; \mathbf{\Theta}^{T+1})}{|\mathbf{v}^t|})$
    Select document $\mathbf{v}^t$ for augmentation in task $T+1$:
    if $\mathrm{PPL}(\mathbf{v}^t; \mathbf{\Theta}^{T+1}) \leq \mathrm{PPL}^{T+1}$ **then**
      Document selected: $\Omega_{aug}^{T+1} \leftarrow \Omega_{aug}^{T+1} \cup (\mathbf{v}^t, t)$
    end if
  end for
  return Ω_{aug}^{T+1}
end function

- **Step 2** *Selective Co-training*

$$\Delta_{SAL} = \sum_{(\boldsymbol{v}^t, t) \in \Omega_{aug}^{T+1}} \lambda_{SAL}^{t} \mathcal{L}(\boldsymbol{v}^t;\ \Theta^{T+1})$$

$$\mathcal{L}(\Omega^{T+1};\ \Theta^{T+1}) = \sum_{\boldsymbol{v} \in \Omega^{T+1}} \mathcal{L}(\boldsymbol{v};\ \Theta^{T+1}) + \Delta_{SAL}$$

- Overall loss in LNTM framework:

$$\mathcal{L}(\Omega^{T+1};\ \Theta^{T+1}) = \sum_{\boldsymbol{v} \in \Omega^{T+1}} \mathcal{L}(\boldsymbol{v};\ \Theta^{T+1}) + \Delta_{TR} + \Delta_{SAL}$$

Neural Topic Modeling with Continual Lifelong Learning

Experiments and Analysis

← Scores on historical data incurring Catastrophic Forgetting → | Scores with Lifelong Knowledge Transfer

	AGnews PPL	AGnews P@0.02	TMN PPL	TMN P@0.02	R21578 PPL	R21578 P@0.02	20NS PPL	20NS P@0.02	R21578title (Sparse Data) PPL	P@5	P@10	P@0.02	COH	r-time (second)
LNTM + EmbTF + TR + SAL	533	0.789	699	0.648	438	0.721	531	0.251	194	**0.828**	**0.810**	**0.690**	0.747	519
LNTM + EmbTF + TR	550	0.788	703	0.650	444	0.721	532	0.251	203	0.812	0.786	0.676	**0.752**	12.63
LNTM + TR	571	0.787	704	0.649	451	0.722	532	0.251	208	0.810	0.770	0.668	0.742	12.18
LNTM + EmbTF	555	0.784	702	0.650	446	0.722	532	0.251	**183**	0.814	0.790	0.678	0.709	11.42
NTM without Lifelong Learning	454	0.785	584	0.651	311	0.726	470	0.268	192	0.799	0.778	0.657	0.713	10.49

AGnews → ***TMN*** → ***R21578*** → ***20NS*** → ***R21578title*** (Sparse Data)

Lifelong Neural Topic Modeling over a stream of document collections →