World Wide Web (2022) 25:219–238
https://doi.org/10.1007/s11280-021-00984-2

Lifelong topic modeling with knowledge-enhanced adversarial network

Xuewen Zhang[1] · **Yanghui Rao**[1] · **Qing Li**[2]

Received: 20 March 2021 / Revised: 24 November 2021 / Accepted: 29 November 2021 /
Published online: 23 December 2021

Abstract
Lifelong topic modeling has attracted much attention in natural language processing (NLP), since it can accumulate knowledge learned from past for the future task. However, the existing lifelong topic models often require complex derivation or only utilize part of the context information. In this study, we propose a knowledge-enhanced adversarial neural topic model (KATM) and extend it to LKATM for lifelong topic modeling. KATM employs a knowledge extractor to encourage the generator to learn interpretable document representations and retrieve knowledge from the generated documents. LKATM incorporates knowledge from the previous trained KATM into the current model to learn from prior models without catastrophic forgetting. Experiments on four benchmark text streams validate the effectiveness of our KATM and LKATM in topic discovery and document classification.

Keywords Neural topic modeling · Lifelong learning · Knowledge distillation

1 Introduction

Learning is often considered as a lifelong process of requiring knowledge and mastering new skills throughout human life. To accumulate knowledge from past and meanwhile avoiding catastrophic forgetting [36], lifelong learning has been studied in a wide range of machine learning tasks [4, 17, 50].

One-shot topic models, such as the latent Dirichlet allocation (LDA) [3], DocNADE [31], the adversarial topic model (ATM) [55], and the bidirectional adversarial topic model (BAT) [54] have shown remarkable success in exploring semantic patterns from

✉ Yanghui Rao
raoyangh@mail.sysu.edu.cn

Xuewen Zhang
zhangxw53@mial2.sysu.edu.cn

Qing Li
csqli@comp.polyu.edu.hk

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

[2] Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

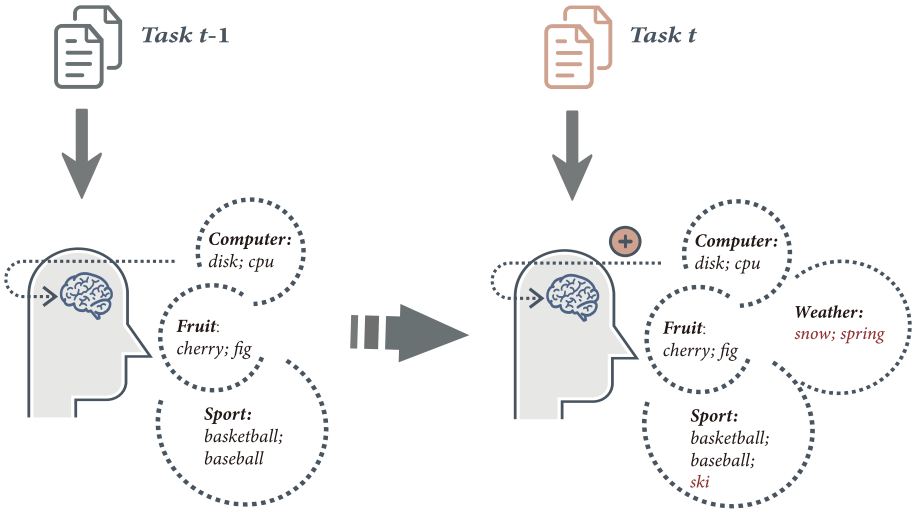

Fig. 1 An example of topic words learning in a lifelong process

a static document collection. Yet the lack of guidance from prior knowledge limits the performance of above methods on text streams. Early efforts have demonstrated how lifelong topic models could be incrementally learned for streaming data [19, 57], but they take the probabilistic perspective to estimate parameters and often involve complex derivation. Recently, a lifelong neural topic model named LNTM [16] is developed based on DocNADE [31] with more flexible training schemes than probabilistic models. However, it only considers the words appeared before the target word while ignores the following words in the sequence [15]. Besides, topic models based on NADE [2], including DocNADE and iDocNADE [15], do not consider the relationship between topics since they are trained in the document-word space. As mentioned in [54], the relationship between topics is useful for improving the model performance on topic coherence and downstream tasks. Therefore, a more generic lifelong neural topic model that can enable continual learning using comprehensive and topic relationship information is valuable.

In this paper, we develop a knowledge-enhanced adversarial neural topic model (KATM) and extend it to LKATM by knowledge distillation and data augmentation. Adversarial neural topic models [22, 54, 55] use a generator network to capture the semantic structure of documents through adversarial training, which overcomes the limitation of complex derivation in probabilistic models and unable to generate coherent topic words in variational auto-encoder (VAE) based neural topic models. To keep the memory of learning previous tasks, we further transfer prior topic information in the current task by knowledge distillation [5]. Figure 1 presents an example of learning topic words in a lifelong process. Suppose we have learned the representative words of three topics (computer, fruit, and sport) from task $t-1$. Given a new task t, we expand the topic words based on previous results and learn a new coherent topic. To achieve this, the main challenges are: (1) how to extract knowledge from the current topic model; (2) how to exploit useful semantic patterns from past models by modeling the topic relationships; (3) how to avoid or minimize catastrophic forgetting of prior topic knowledge. In light of these considerations, we summarize the main contributions of this work as follows:

- We develop the KATM by training a knowledge extractor to retrieve semantic patterns of documents generated from the generator network. This enables our model to extract topic knowledge from the generator and encourages it to learn more interpretable document representations.
- We propose the LKATM to incorporate semantic patterns from previous trained models into the current model and utilize data augmentation to avoid the conflicts caused by the inconsistent output of different models. To the best of our knowledge, we are the first to develop a lifelong neural topic model based on adversarial networks, in addition to utilize the principle of knowledge distillation for lifelong neural topic modeling.

We evaluate the effectiveness of our KATM and LKATM on four real-world text streams. Experimental results demonstrate that the coherence and uniqueness of topics generated by our models are improved significantly when compared with state-of-the-art approaches. The quality of document representations from different models has also been tested on document classification.

2 Related work

In this section, we briefly introduce lifelong machine learning, neural topic modeling, and knowledge distillation which are related to our work.

2.1 Lifelong machine learning

Lifelong machine learning is capable of training a model from data streams. It aims to integrate the current knowledge into the model without catastrophic forgetting over time [44]. The existing lifelong machine learning studies mainly focus on the following research directions: (1) Dynamic architecture based methods [8, 35, 50] which expand model architectures for new tasks to avoid losing the previous learned knowledge, e.g., re-training with an additional number of neurons or network layers. While the methods of introducing new neurons and network layers alleviate the catastrophic forgetting issue in nonstationary environments [40], they do not resemble biologically plausible mechanisms [44] and may be inapplicable to natural language processing models with fixed neurons. (2) Lifelong machine learning with auxiliary data [5, 17, 47, 48] which restores a few examples in previous tasks and incorporates them into the current task to tackle catastrophic forgetting. This is similar to humans who review previous tasks to acquire knowledge. By training with the same data sampled from each task, a method can study the shared high-level representations of streaming data. Learning with auxiliary data has been widely studied for over two decades and still used nowadays because of its effectiveness. (3) Parameter consolidation methods [16, 53] which constraint on the update of the neural weights. Strategies of emphasizing important parameters from previous tasks have been proposed in [16, 30, 60], e.g., introducing a quadratic penalty on the difference between the parameters for prior and new tasks. However, these methods may lead to calculation issues if the neural architectures become very large. On the other hand, Donahue et al. [9] attempt to prevent significant changes in the network parameters when training with new data by reducing the learning rate. Besides, a regularization term related to the prior loss [32] is proposed to mitigate catastrophic forgetting. Unfortunately, its effectiveness is highly affected by the performance of previous models. In summary, parameter consolidation methods provide a

way to learn continual tasks under certain conditions [44] while they are still worthy of further researches. Different from the above lifelong machine learning methods, our approach aims to minimize the difference of knowledge that is extracted from tasks with auxiliary information over data streams. Specifically, given a new task, the current model learns a soft target extracted from previous models to minimize catastrophic forgetting in the lifelong process.

2.2 Neural topic modeling

Topic modeling has been widely used in text mining, including document clustering, information recommendation, and information retrieval [11, 23, 26]. Traditional topic models rely on approximate approaches (e.g., variational inference and Gibbs sampling) to estimate parameters [3, 25]. However, variational inference often involves complex derivation and Gibbs sampling requires high computational costs. To address these weaknesses, VAE and neural variational inference (NVI) [41] are used as the frameworks of several preliminary neural topic models [37, 38, 51] due to their flexible and fast parameter inference.

With the rapid development of generative adversarial net (GAN) [14], there is a new direction to discover topics based on GAN. For instance, ATM [55] is proposed by using Dirichlet priors for latent topics instead of multivariate Gaussian priors or logistic-normal priors. This model aims to train a generator to learn the mapping from the document-topic distribution to the document-word distribution. Inspired by bidirectional adversarial training, BAT [54] builds an encoder to capture real topic distributions combined with fake distributions from the generator. To handle labeled documents, a cycle-consistent adversarial topic model [22] is proposed. Apart from the above methods, the adversarial-neural event model [56] is proposed for extracting the structured representations of open-domain events. To address the lack of data representations in the topic space and the limitation of spending a lot of time to manually label useful topics, a reward function and a topic predictor are integrated into GAN [12]. In our approach, a knowledge extractor is added into GAN, which aims to encourage the generator to learn more interpretable and meaningful representations, by minimizing the difference between the generator input and the knowledge extractor output.

2.3 Knowledge-enhanced NLP methods

With the development of deep learning technologies, the input text alone contains limited knowledge to support models producing satisfactory output. Incorporating knowledge into NLP models becomes a promising direction in both academia and industry [59]. Recently, developing specialized architectures is widely studied to process knowledge, including attention network based methods [7, 13, 18, 46], graph neural network based methods [61, 62], and memory network based methods [34, 58]. Knowledge-enhanced learning is agnostic to the model architecture and can be combined with various architectures. However, the sources of knowledge should not be limited to a single network structure, dictionary, and table [59]. The reason is that knowledge transferring by learning from multi-domain sources can discover knowledge more broadly and meanwhile improve the knowledge generation process.

Knowledge distillation is an effective solution for knowledge transferring, by using the predicted distributions of a teacher model as soft targets to train a less-parameterized student model [20]. Recent efforts have demonstrated how the refined soft predictions could

improve the generation of student model as compared with hard labels [28, 33]. Furthermore, the flexible methods have extended to scenarios where all student models distill knowledge without a pre-trained teacher model by learning from peers' predictions [6].

In the field of topic modeling, BERT-based auto-encoder teacher model [21] combines the advantages of probabilistic topic models and pre-trained transformers by mapping documents through a standard bag-of-words representation and a teacher model. Unlike the above model, our LKATM directly takes the model trained from previous task as the teacher model to generate the current document-topic distribution better.

3 Methodology

In this section, we firstly describe the task of lifelong topic modeling. Then, we introduce the proposed KATM. Finally, we extend KATM to LKATM with knowledge distillation and data augmentation for lifelong topic modeling.

3.1 Problem formulation

Consider a stream of documents $\Omega = \{\Omega^1, \Omega^2, \Omega^3, ...\}$ accumulated over lifetime. During the training of the t^{th} task, there are document collections of $\mathbb{D}^t$ paired instances $\{(\boldsymbol{d}_{\boldsymbol{r}}^t, \boldsymbol{\theta}^t) | \boldsymbol{d}_{\boldsymbol{r}}^t \in \boldsymbol{d}_{\boldsymbol{r}}, \boldsymbol{\theta}^t \in \boldsymbol{\theta}\}_{t=1}^{+\infty}$ where $\boldsymbol{d}_{\boldsymbol{r}}^t$ denotes the set of real documents and $\boldsymbol{\theta}^t$ denotes the topic distribution to generate the corresponding fake document $\boldsymbol{d}_{\boldsymbol{f}}^t$. For the t^{th} trained model M^t, the goal is to generate $\boldsymbol{d}_{\boldsymbol{f}}^t \leftarrow \boldsymbol{\theta}^t$ as similar as $\boldsymbol{d}_{\boldsymbol{r}}^t$, without forgetting how to generate documents of previous tasks $\boldsymbol{d}_{\boldsymbol{f}}^j \leftarrow \boldsymbol{\theta}^j$ where $j = (1, 2, ..., t-1)$.

Inspired by knowledge distillation, a student model is trained by the predicted soft distribution from a teacher model, in which, we treat the current model M^t as the student model and M^{t-1} as the teacher model. Knowledge distillation is used to extract valuable information from M^{t-1} to M^t by encouraging these two models to produce similar output or patterns with the same data as input. In addition, the corpus $C = \{C^1, C^2, C^3, ...\}$ is augmented each time followed by training a new task to accumulate knowledge.

3.2 KATM: Knowledge-enhanced adversarial neural topic model

We here present our KATM, which aims to encourage the generator to learn more interpretable and meaningful document representations. We accomplish it by minimizing the difference between the sampled document-topic distribution $\boldsymbol{\theta}$ and the generated document-topic distribution $\widetilde{\boldsymbol{\theta}}$. As shown in Figure 2, KATM contains four components: a real document set, a generator G, a discriminator D, and a knowledge extractor E. The generator contains a K-dimensional document-topic distribution layer, an S-dimensional representation layer, and a V-dimensional document-word distribution layer. The discriminator consists of a V-dimensional document-word distribution layer, an S-dimensional representation layer, and an output layer. The knowledge extractor included in discriminator contains a K-dimensional document-topic distribution layer, which generates the document-topic distribution $\widetilde{\boldsymbol{\theta}}$ by softmax normalization.

Following ATM [55], we train generator G to obtain a document-word distribution by transforming a K-dimensional noise variable $\boldsymbol{\theta} \sim Dir(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ into a V-dimensional sample $\boldsymbol{d}_f$, where $\boldsymbol{\alpha}$ is the hyperparameter of Dirichlet distribution. The generator is

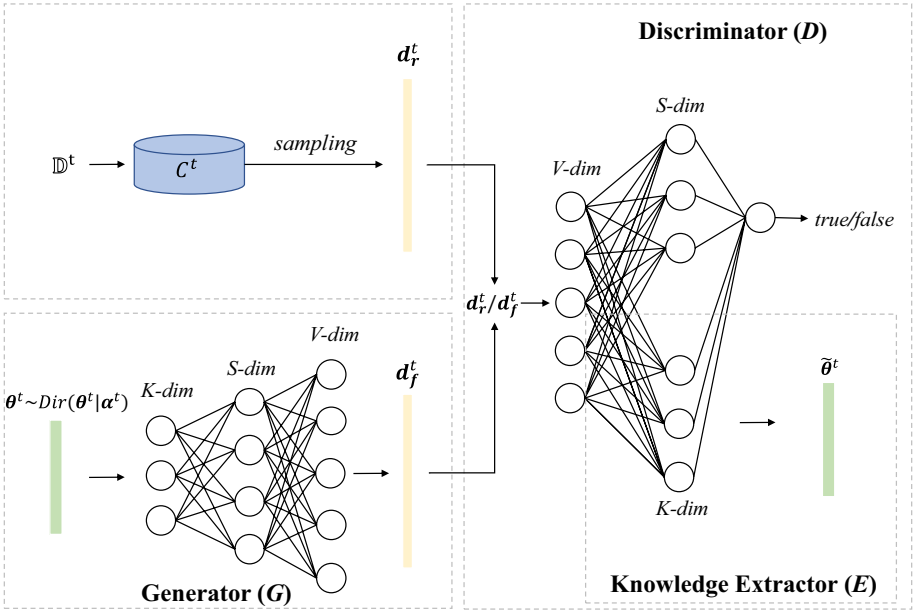

Fig. 2 The framework of KATM on the t^{th} task, i.e., M^t

guided by an adversarial discriminator D which aims to distinguish the fake document-word distribution $\boldsymbol{d_f}$ from the true document-word distribution $\boldsymbol{d_r}$. The real documents in the corpus are represented by TFIDF, noted as $\mathbb{P}_r$. The real distributions can be viewed as random samples drawn from $\mathbb{P}_r$. Formally, the adversarial loss is given by $\min_G \max_D V(D,G) = E_{\boldsymbol{d_r}\sim\mathbb{P}_r}[\log D(\boldsymbol{d_r})] + E_{\boldsymbol{\theta}\sim Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})}[\log(1 - D(G(\boldsymbol{\theta})))]$.

Note that the noise vector fed into the generator is pre-determined and fixed. Different from [54, 55], we propose to take the input vector as a latent code and train it as the target rather than simply take it as a noise vector. In this way, our model not only gets the semantic feature of documents better through learning prior knowledge, but also infers the document-topic distribution explicitly. Particularly, we develop a knowledge extractor to capture the document-topic distribution from each generated document. As shown in the bottom-right part of Figure 2, the knowledge extractor is a K-dimensional single-layer neural network included in discriminator. As part of the discriminator's embedding layer, it takes the fake document $\boldsymbol{d_f}$ as input and outputs the topic distributions $\widetilde{\boldsymbol{\theta}}$ by softmax normalization. We use the weight of knowledge extractor, i.e., a $K \times V$-dimensional matrix, as the topic-word distribution.

Suppose generator G could generate documents the same as sampling from the corpus and knowledge extractor E could retrieve the semantics of fake documents, the difference between prior document-topic distributions $\boldsymbol{\theta}$ and output $\widetilde{\boldsymbol{\theta}}$ should be small. The adversarial loss mentioned above encourages the generator to generate documents matching the data distribution in the corpus, and meanwhile the knowledge layer loss promotes the generator to construct a more explainable document containing some given semantic information. Specifically, the Kullback-Leibler (KL) divergence between $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ is used to define the aforementioned difference, as follows:

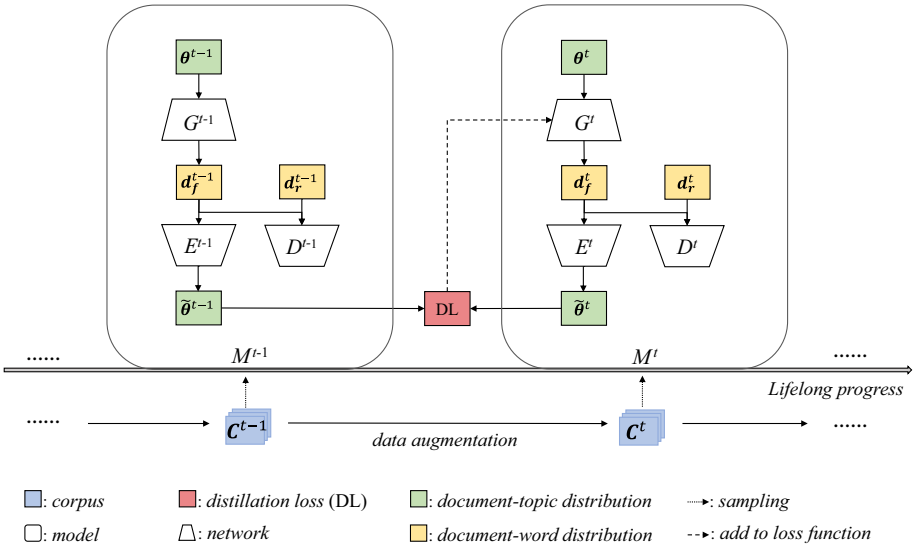

Fig. 3 Overview of LKATM

$$\mathcal{L}_K = \sum_i \boldsymbol{\theta}_i \log \frac{\boldsymbol{\theta}_i}{\widetilde{\boldsymbol{\theta}}_i}. \tag{1}$$

Finally, KATM's loss function is defined as the following formula with a regularization term of KL divergence between $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$:

$$\min_{G,E} \max_{D} V_{KATM}(D, G, E) = V(D, G) + \lambda_k \mathcal{L}_K, \tag{2}$$

where λ_k is a hyperparameter.

3.3 LKATM: Lifelong knowledge-enhanced adversarial neural topic model

Knowledge distillation The simplified form of knowledge distillation is defined as follows: a student model is trained by a soft target distribution which is produced by a teacher model with a user-specified temperature. Given the teacher model's output of the last fully connected layer g_i and temperature T, the soft output θ_i is defined by:

$$\theta_i = \frac{\exp\left(g_i/T\right)}{\sum_j \exp\left(g_j/T\right)}. \tag{3}$$

Then, knowledge is transferred by combining the student model's predicted distribution, which is produced using the same temperature T in such a model's soft output, with the teacher model's distribution $\boldsymbol{\theta}$. A higher T means a softer distribution. Based on KATM, Figure 3 presents the framework of LKATM that enables topical knowledge transfer from different domains without catastrophic forgetting. It can also be understood as distilling document-topic distributions generated from previous tasks. As mentioned earlier, KATM outputs $\widetilde{\boldsymbol{\theta}}$, i.e., a document-topic distribution from the knowledge extractor. Given the same

noise document-topic distribution $\boldsymbol{\theta}$, the models M^t and M^{t-1} are encouraged to generate the same output. In our approach, we define the loss for knowledge distillation as follows:

$$\mathcal{L}_{DL} = KL(\widetilde{\boldsymbol{\theta}}^{t-1}, \widetilde{\boldsymbol{\theta}}^{t}) = \sum_i \widetilde{\boldsymbol{\theta}}_i^{t-1} \log \frac{\widetilde{\boldsymbol{\theta}}_i^{t-1}}{\widetilde{\boldsymbol{\theta}}_i^{t}}. \tag{4}$$

The overall loss of the t^{th} task is given below:

$$\min_{G^t,E^t} \max_{D^t} V_{LKATM}(D^t, G^t, E^t) = V(D^t, G^t) + \lambda_k \mathcal{L}_K + T^2 \mathcal{L}_{DL}, \tag{5}$$

where $\mathcal{L}_{DL}$ is multiplied by T^2 to ensure that the relative contribution of distillation term remains roughly unchanged if the temperature is changed while experimenting with meta-parameters [20].

Algorithm 1 Algorithm of LKATM

Input: $\mathbb{P}_r^t, M^{t-1}, C^{t-1}, n_d^t, m^t, c^t, \alpha^t, \lambda_k^t, \Theta^t, Z, T$
Output: M^t

Initialize D^t parameters ω_d^t, G^t parameters ω_g^t and E^t parameters ω_e^t
while ω_d^t, ω_g^t, and ω_e^t have not converged **do**
  for $i = 1, ..., n_d^t$ **do**
    for $j = 1, ..., m^t$ **do**
      sample $\boldsymbol{d_r^t} \sim \mathbb{P}_r^t$
      sample $\boldsymbol{\theta}^t \sim Dir\left(\boldsymbol{\theta}^t \mid \boldsymbol{\alpha}^t\right)$
      $\boldsymbol{d_f^t} \leftarrow G(\boldsymbol{\theta}^t), \widetilde{\boldsymbol{\theta}}^t \leftarrow E(\boldsymbol{d_f^t})$
      compute $\mathcal{L}_k^j(\boldsymbol{\theta}^t, \widetilde{\boldsymbol{\theta}}^t)$ by Eq. (1)
      compute $\mathcal{L}_{DL}^j(\widetilde{\boldsymbol{\theta}}^{t-1}, \widetilde{\boldsymbol{\theta}}^t)$ by Eq. (4)
      $\mathcal{L}_d^j \leftarrow D(\boldsymbol{d_f^t}) - D(\boldsymbol{d_r^t})$
      $\mathcal{L}_g^j \leftarrow -D(\boldsymbol{d_f^t}) + T^2 \mathcal{L}_{DL}^j$
    end for
    $\omega_d^t \leftarrow Adam(\nabla_{\omega_d^t} \frac{1}{m^t} \sum_{j=1}^{m^t} \mathcal{L}_d^j, \omega_d^t, \Theta^t)$
    $\omega_d^t \leftarrow clip(\omega_d^t, -c^t, c^t)$
  end for
  $\omega_g^t \leftarrow Adam(\nabla_{\omega_g^t} \frac{1}{m^t} \sum_{j=1}^{m^t} \mathcal{L}_g^j, \omega_g^t, \Theta^t)$
  $\omega_e^t \leftarrow Adam(\nabla_{\omega_e^t} \frac{1}{m^t} \sum_{j=1}^{m^t} \lambda_k^t \mathcal{L}_k^j, \omega_e^t, \Theta^t)$
  $\omega_e^t \leftarrow clip(\omega_e^t, -c^t, c^t)$
end while
Update C^t

Data augmentation The performance of deep learning models largely depends on the amount of training data [10]. Data augmentation attempts to manipulate data for training to improve the model's generalization ability. Note that (5) contains conflicted objectives. The first and second items encourage inputs to fit model M^t, while the third item encourages M^t to generate the same output as that of model M^{t-1}. The conflicts make it difficult for the model to learn topics efficiently.

To tackle the above problem, we propose to use data augmentation by adding top N real documents in Ω^{t-1} evaluated by the performance on discriminator D^{t-1} into the corpus C^t. The use of data augmentation can remove these conflicts. In addition, the vocabulary

Table 1 The statistics of datasets

Datasets	#Documets	#Words
AGnews	1,879	821
TMN	8,275	1,270
R21578	7,265	4,419
20NS	8,775	6,128
grolier	28,938	3,000

dimension of ground truth data will change across domains. The dimensional mismatch problem occurs when the model is training, so it is necessary to update C^t each time, which is a common operation of data augmentation.

The overall algorithm is shown in Algorithm 1. For the t^{th} task, we use n_d^t to denote the number of discriminator's training iterations per generation iteration. Furthermore, m^t is the batch size, c^t is the clipping parameter, and λ_k^t represents the weight of knowledge extractor.

4 Experiments

In this section, we evaluate our KATM and LKATM by answering the following questions.

- Q1. Does KATM effectively minimize the difference between prior and generated document-topic distributions? (Section 4.1)
- Q2. Does knowledge extractor learn better topic-word distributions than generator? (Section 4.1)
- Q3. How does KATM perform when compared with other one-shot neural topic models? (Sections 4.2 and 4.3)
- Q4. How does LKATM perform when compared with the state-of-the-art lifelong neural topic model? (Sections 4.2 and 4.3)
- Q5. How does temperature affect LKATM's performance? (Section 4.4)
- Q6. How does LNTM perform in a downstream task? (Section 4.5)

Datasets Following [16], we use four real-word datasets for training: AGnews, Tag My News (TMN), Reuters 21578 corpus (R21578), and 20NewsGroups corpus (20NS). Note that non UTF-8 characters and stop words are eliminated. All test datasets (i.e., 20NSshort, TMNtitle, and R21578title) in [16] are used to measure the model performance over short text. Besides these short-text test datasets, we also employ a long-text test dataset, i.e., grolier[1] to perform lifelong topic modeling. The statistics of datasets are shown in Table 1. Similar to [16], we construct the following data streams for evaluation:

- *AGnews* → *TMN* → *R*21578 → 20*NS* → 20*NSshort*
- *AGnews* → *TMN* → *R*21578 → 20*NS* → *TMNtitle*
- *AGnews* → *TMN* → *R*21578 → 20*NS* → *R*21578*title*

[1] https://cs.nyu.edu/~roweis/data.html

Table 2 Characteristics of baselines and our models

Model	Methodology	Type	Use external knowledge?
NVDM	NVI [41] based	One-shot	No
NVLDA	NVI [41] baesd	One-shot	No
DocNADE	NADE [31] based	One-shot	No
iDocNADE	NADE [31] based	One-shot	Yes
ATM	GAN [14] based	One-shot	No
BAT	GAN [14] based	One-shot	No
SCH. + BAT	VAE [51] based	One-shot	Yes
LNTM	NADE [31] based	Lifelong	Yes
KATM (ours)	GAN [14] based	One-shot	No
LKATM (ours)	GAN [14] based	Lifelong	No

– *AGnews* → *TMN* → *R*21578 → 20*NS* → *grolier*

Taking the second data stream as an example, we train sequentially on AGnews, TMN, R21578, and 20NS in a lifelong process. Based on the prior models, TMNtitle is adopted as the current dataset to demonstrate whether catastrophic forgetting is avoided. After training on TMNtitle, the current model is used to generate topics and applied for downstream tasks. Specifically, our model includes a generator, a discriminator, and a knowledge extractor. As topics are generated from the current knowledge extractor, our experiments, i.e., topic quality comparison and document classification, are mainly carried out on it. To perform the task of document classification in the lifelong process, we first get the trained model mentioned above, and then input the document-word distributions which are converted from the current dataset into the knowledge extractor and take the outputs as document-topic distributions. More details will be introduced in Section 4.5. For clarity, the descriptions of all datasets are given below:

1. AGnews: a data collection provided by ComeToMyHead for research purposes in text mining, information retrieval, and so forth.
2. TMN: a news dataset labelled with 7 categories. Each news story contains a title and a description.
3. R21578: a collection of new stories from the natural language toolkit (NLTK)[2]. NLTK is a suite of open source Python modules, data sets, and tutorials.
4. 20NS: a collection of news stories partitioned across 20 newsgroups.
5. grolier: the Grolier multimedia encyclopedia articles. Its content covers almost all the fields in the world, such as sports, economics, and politics.
6. TMNtitle: titles of the TMN dataset.
7. R21578title: titles of the R21578 corpus.
8. 20NSshort: documents from 20NS with document size (i.e., the number of words in a document) less than 20.

[2] http://www.nltk.org/data.html

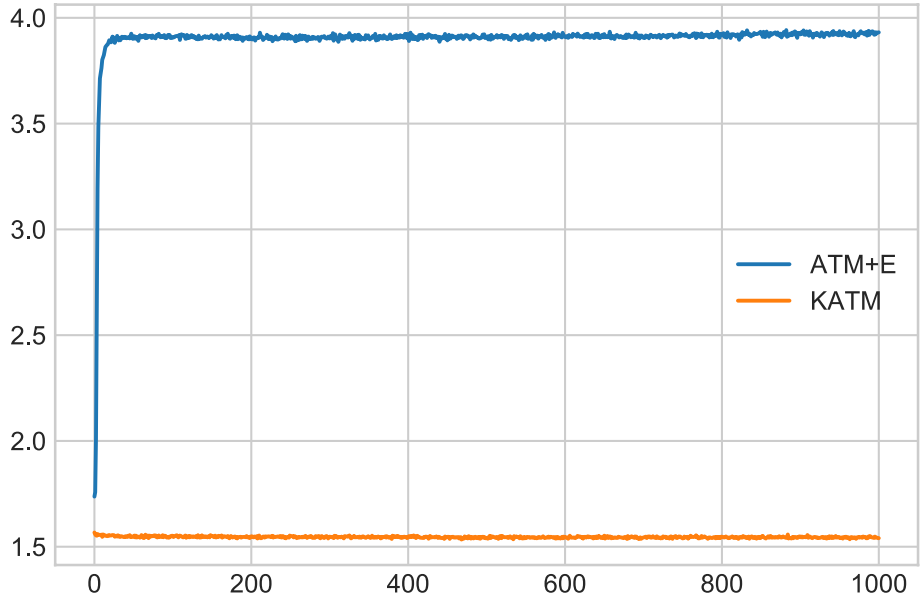

Fig. 4 Divergence of ATM+E and KATM over training iterations

Baselines We adopt the following models for comparison: NVDM [38], NVLDA [51], DocNADE [31], iDocNADE [15], ATM [55], BAT [54], SCH. + BAT [21], and LNTM [16]. For completeness, we present the characteristics of these baselines and our models in Table 2.

Network architecture For generator, discriminator, and knowledge extractor in KATM and LKATM, we use feed-forward neural networks with ReLU activation [42] and batch normalization (BN) [24]. The detailed transformations of generator are: [Linear(K, S) → ReLU → BN → Linear(S, V) → Softmax], those of discriminator are: [BN → Linear($V, S + K$) → ReLU → Linear($S + K, 1$)], and those of knowledge extractor are: [BN → Linear(V, K) → Softmax]. In the above, Linear() denotes a linear transformation.

In our experiments, we set the hyperparameters of KATM and LKATM as follows: $n_d = 5, m = 64, c = 0.01, \lambda_k = 1, T = 3$, and $S = 150$. We update model parameters using Adam [29] with $\beta_1 = 0.5, \beta_2 = 0.999$, and $\epsilon = 10^{-4}$.

4.1 Effectiveness of the knowledge extractor

As mentioned earlier, the knowledge extractor is trained to refine topic distributions of the generated documents the same as the sampled document-topic distributions. To evaluate whether the sampled document-topic distributions are similar as the generated document-topic distributions trained by the proposed method, we train KATM on the 20NS dataset with 50 topics. The result in Figure 4 indicates that the divergence $\mathcal{L}_k$ is maintained at about 1.50.

For comparison, we also train ATM [55] with an auxiliary knowledge extractor E when the generator is not explicitly encouraged to minimize the divergence between prior and

generated document- topic distributions. The result shows that the divergence in ATM+E quickly increases to $\mathcal{L}_k \approx 3.90$. This indicates that in ATM+E, there is no guarantee that the generator could make use of document-topic distributions to generate documents with enough semantic information.

In addition, to explore whether the knowledge extractor can generate topic-word distributions effectively, we use another adversarial network to generate the word distribution of each topic from the generator instead of the knowledge extractor, which is named as KATM_G. We use four widely-adopted topic coherence metrics, i.e., C_V [49], C_A [1], NPMI [1], and UMass [39] to evaluate the performance of different models. A higher topic coherence value means more understandable topics are extracted. All coherence values are calculated by the Palmetto library[3] over top 10 words of 50 topics according to the generated topic-word distribution. The result is shown in Table 3, which indicates that topic words generated by the knowledge extractor in our KATM achieves better performance than KATM_G on AGnews, TMN, R21578, and 20NS. This validates that the knowledge extractor in adversarial neural topic models is useful to capture coherent topics.

4.2 Topic coherence comparison

In this task, we evaluate the performance of the proposed models and baselines using topic coherence metrics mentioned above. The numbers of topics are set to 20, 30, 50, 75, and 100, except for 50 and 100 in iDocNADE since it represents a document by summing the vectors of its words through Glove embeddings [45]. The averaged coherence scores are calculated as the final results, as shown in Table 3. These results indicate that the performance of KATM is better than others (except for iDocNADE, which performs better in UMass score on 20NS). Furthermore, LKATM maintains competitive topic coherence scores in the lifelong process, even better than KATM on TMN and 20NS. It validates that LKATM can effectively avoid catastrophic forgetting and learn from past models to obtain high quality topics.

In addition, we compare the average topic coherence scores of our LKATM with the existing lifelong neural topic model LNTM. The detailed topic coherence scores are shown in Table 4. Since LNTM represents a document by summing the word vectors through Glove embeddings [45], the topic numbers are set to 50 and 100 included in the pre-trained Glove model to calculate topic coherence. Each value is calculated by averaging coherence scores over top 10 words. We highlight the best topic coherence value on each metric by boldface. Among all the metrics, LKATM achieves the best performance on training datasets (AGnews, TMN, R21578, and 20NS), and also better on testing datasets (20NSshort, R21578title, TMNtitle, and grolier).

As an illustration, Table 5 presents top 10 words of 4 representative topics extracted by LNTM and LKATM. The result shows that the proposed LKATM can generate more coherent topics.

4.3 Topic uniqueness comparison

As mentioned in [43], neural topic models tend to generate high coherence scores but identical topics to minimize loss. It is also important to generate topics which are diverse instead of repetitive. Thus, we compute topic uniqueness (TU) scores proposed

[3] http://aksw.org/Projects/Palmetto.html

Table 3 Average topic coherence scores on AGnews, TMN, R21578, and 20NS with topic number setting as [20, 30, 50, 75, 100]. Given a dataset, the best value on each metric is highlighted by boldface

Dataset	Model	C_V	C_A	NPMI	UMass
AGnews	NVDM	0.3743	0.1295	–0.0773	-4.2274
	NVLDA	0.3745	0.1350	–0.0677	–4.2298
	DocNADE	0.4042	0.1177	–0.1058	–5.6555
	iDocNADE	0.4081	0.1103	–0.1155	–6.6869
	ATM	0.3716	0.1214	–0.0703	–3.7029
	BAT	0.3733	0.1191	–0.0810	–4.1172
	SCH.+BAT.	0.3306	0.1278	–0.0759	–3.4655
	KATM_G	0.3531	0.1191	–0.0662	–3.9672
	KATM (ours)	**0.4213**	**0.1405**	**–0.0620**	**–3.3986**
TMN	NVDM	0.3776	0.1321	–0.0717	–4.2060
	NVLDA	0.3684	0.1301	–0.0604	–3.1052
	DocNADE	0.3444	0.1243	–0.0577	–2.9073
	iDocNADE	0.3653	0.1145	–0.0557	–3.3474
	ATM	0.3877	0.1420	–0.0557	–4.4915
	BAT	0.3772	0.1222	–0.0935	–4.1610
	SCH.+BAT.	0.3940	0.1240	–0.0882	–4.2689
	KATM_G	0.4044	0.1191	–0.0618	–3.7545
	KATM (ours)	**0.4168**	0.1410	–0.0665	–3.2082
	LKATM (ours)	0.4094	**0.1435**	**–0.0542**	**–1.7572**
R21578	NVDM	0.3924	0.1313	–0.1170	–4.9407
	NVLDA	0.3817	0.1391	–0.0656	–4.5007
	DocNADE	0.3748	0.1296	–0.0810	–4.0862
	iDocNADE	0.3705	0.1231	–0.0834	–4.5417
	ATM	0.3877	0.1420	–0.0757	–4.4915
	BAT	0.3624	0.1341	–0.0694	–3.5385
	SCH.+BAT.	0.4001	0.1210	–0.0959	–5.2718
	KATM_G	0.3868	0.1191	–0.0801	–4.6143
	KATM (ours)	**0.4336**	**0.1423**	**–0.0593**	–4.0888
	LKATM (ours)	0.4305	0.1403	–0.0895	**–2.2551**
20NS	NVDM	0.3905	0.1264	–0.1160	–5.5530
	NVLDA	0.3821	0.1281	–0.0614	–4.2185
	DocNADE	0.3677	0.1299	–0.0622	–3.9732
	iDocNADE	0.3535	0.1180	–0.0641	**–2.6284**
	ATM	0.3976	0.1360	–0.0642	–4.6057
	BAT	0.3807	0.1246	–0.1124	–4.7292
	SCH.+BAT.	0.4180	0.1179	–0.0937	–5.8701
	KATM_G	0.3752	0.1191	–0.0683	–4.0995
	KATM (ours)	0.4452	**0.1770**	–0.0651	–4.6973
	LKATM (ours)	**0.4474**	0.1406	**–0.0587**	–4.4525

in [43] to estimate the discrimination of topics. Given a set of top-n representative words from each of the K topics, the TU score for topic k is inversely proportional to the number of times each of top-n word is repeated in the set. And the average TU is computed by $\text{TU} = \frac{1}{K}\sum_{k=1}^{K}\text{TU}(k)$. The range of TU value is between $\frac{1}{K}$ and 1. A higher TU means the produced K topics are more diverse.

Table 4 Average topic coherence scores on AGnews, TMN, R21578, 20NS, 20NSshort, R21578title, TMNtitle, and grolier with topic number setting as [50, 100]. Given a dataset, the best value on each metric is highlighted by boldface

Dataset	Model	C_V	C_A	NPMI	UMass
AGnews	LNTM	0.3914	0.1162	–0.0984	–5.7633
	LKATM	**0.4119**	**0.1405**	**–0.0670**	**–3.3986**
TMN	LNTM	0.4007	0.1452	–0.1114	–4.2281
	LKATM	**0.4094**	**0.1498**	**–0.0542**	**–1.7572**
R21578	LNTM	0.4257	0.1213	–0.1148	–5.5962
	LKATM	**0.4305**	**0.1425**	**–0.0895**	**–2.2551**
20NS	LNTM	0.4181	0.1359	–0.1491	–5.5393
	LKATM	**0.4409**	**0.1411**	**–0.0860**	**–4.3327**
20NSshort	LNTM	0.4193	0.1209	–0.1191	–5.5101
	LKATM	**0.4631**	**0.1480**	**–0.0886**	**–4.0132**
R21578title	LNTM	0.3939	0.1201	–0.1137	–6.1013
	LKATM	**0.4054**	**0.1454**	**–0.0654**	**–4.5394**
TMNtitle	LNTM	0.4048	0.1215	–0.1098	–6.3827
	LKATM	**0.4364**	**0.1294**	**–0.0848**	**–4.9566**
grolier	LNTM	0.3702	0.1454	–0.0614	–3.3951
	LKATM (ours)	**0.4133**	**0.1493**	**–0.0396**	**–3.0601**

Table 5 Top 10 words of 4 representative topics extracted by LNTM and LKATM, where irrelevant words are marked by italics. These 4 topics indicate 'compute', 'political', 'sports', and 'agriculture', respectively

Model	T1	T2	T3	T4
LNTM	window	governor	baseball	agriculture
	application	committee	play	farmers
	controller	*completes*	game	tonnes
	overall	senate	season	commodity
	microsoft	*undisclosed*	*minutes*	quotas
	disk	*disclosed*	bike	grain
	cpu	*note*	prior	water
	church	community	*windows*	*said*
	chip	states	hockey	*organization*
	play	voted	shipping	natural
LKATM (ours)	disk	government	basketball	farmers
	cpu	president	nba	planting
	machine	states	playoffs	corn
	drive	world	baseball	crop
	keyboard	official	winner	agriculture
	system	minister	sports	weather
	windows	law	coach	feed
	font	political	scoring	cotton
	file	unions	hockey	growers
	zip	conference	grace	rain

We compare the TU scores of our KATM with one-shot topic models mentioned above. The result is shown in Table 6, from which we can observe that GAN based methods (i.e., ATM, BAT, and KATM) can generate more diverse topics than other

Table 6 TU scores of one-shot models with 50 and 100 topics, where top 10 words are used for calculation. The best value on each dataset is highlighted by boldface

Model	AGnews	TMN	R21578	20NS
NVDM	0.482	0.430	0.523	0.520
NVLDA	0.288	0.420	0.524	0.545
DocNADE	0.098	0.232	0.329	0.391
iDocNADE	0.100	0.312	0.342	0.411
ATM	0.566	0.436	0.418	0.441
BAT	0.602	0.508	0.660	0.631
KATM (ours)	**0.750**	**0.535**	**0.715**	**0.740**

Table 7 TU scores of lifelong models with 500 and 100 topics, where top 10 words are used for calculation. The best value on each dataset is highlighted by boldface

Model	AGnews	TMN	R21578	20NS	grolier
LNTM	0.100	0.584	0.535	0.508	0.465
LKATM (ours)	**0.825**	**0.588**	**0.634**	**0.602**	**0.541**

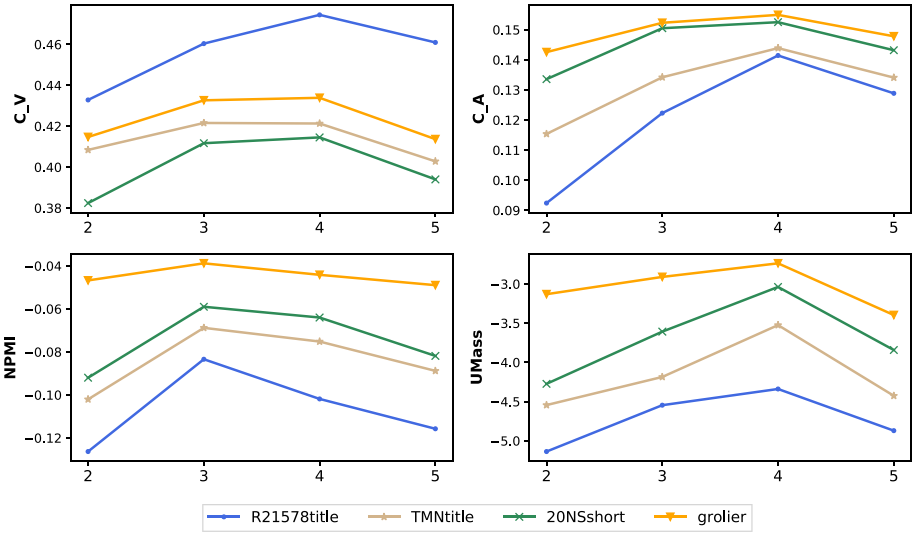

Fig. 5 Topic coherence scores on C_V, C_A, NPMI, and UMass at different temperatures

models. Table 7 presents the TU scores of the proposed lifelong learning method LKATM and LNTM. We observe that LKATM achieves higher TU scores than LNTM across all the datasets, which indicates that LKATM captures more diverse topics.

4.4 Impact of the temperature

To explore how the topic coherence scores vary with respect to different temperatures for our LKATM, we show the topic coherences on four test datasets in Figure 5. In terms of C_V, C_A, and UMass, our LKATM achieves the best performance when the temperature is set to 4. While for NPMI, the best-performing temperature in LKATM is

Table 8 Classification results of DocNADE, iDocNADE, LNTM, and LKATM when combined with the LightGBM classifier. Given a dataset, the best value on each metric is highlighted by boldface

Data Repr.	TMN			20NS		
	MacroF1	MicroF1	AUC	MacroF1	MicroF1	AUC
TFIDF	0.226	0.291	0.621	0.244	0.248	0.720
TFIDF+DocNADE	0.299	0.329	0.662	0.364	0.368	0.789
TFIDF+iDocNADE	0.391	0.402	0.664	0.367	0.371	0.801
TFIDF+LNTM	0.438	0.522	0.771	0.441	0.466	0.791
TFIDF+LKATM (ours)	**0.573**	**0.612**	**0.829**	**0.544**	**0.545**	**0.816**

3. The result indicates that either low or high temperatures will reduce the topic quality. This is because a low temperature may not distill sufficient knowledge, while a high temperature distills too much knowledge to learn the current task. Besides, different datasets present approximately the same trend.

4.5 Application to document classification

Our method is able to generate more coherent topics and potentially interpretable document representations, which can be beneficial to downstream tasks such as document classification. We employ TMN and 20NS datasets and compare the proposed LKATM with DocNADE, iDocNADE, and LNTM in this experiment. For each dataset, we randomly select 80% and 20% data as the training set and the testing set, respectively. LightGBM [27], a highly efficient gradient boosting decision tree, is adopted as the classifier. Particularly, it takes the document-word distribution represented by TFIDF as the input [52]. In our method, we use TFIDF as the knowledge extractor's input and obtain the topic distributions by softmax normalization, that is, we represent each text using the product of TFIDF and the transposed topic-word distributions. To ensure fair comparisons, the same process is performed for all baselines. Table 8 presents the MacroF1, MicroF1, and AUC scores of the LightGBM classifier when using the original TFIDF and the text representations based on DocNADE, iDocNADE, LNTM, and our LKATM. The results indicate that the representation of documents generated by our method is much better to document classification as compared with these baselines.

5 Conclusions

In this work, we proposed a knowledge-enhanced topic model named KATM and a lifelong neural topic model based on KATM (i.e., LKATM) for capturing coherent topics. KATM discovers topics in a document by training a knowledge extractor, which promotes the generator to train more meaningful documents by processing each input vector as a target. LKATM utilizes knowledge distillation and data augmentation to transfer prior topic cues into the current task while avoiding catastrophic forgetting. We empirically demonstrated that the proposed methods achieved better topic coherence and uniqueness than state-of-the-art topic models on various benchmark datasets.

Acknowledgements The research described in this paper was supported by the National Natural Science Foundation of China (61972426), Guangdong Basic and Applied Basic Research Foundation (2020A1515010536), the Hong Kong Research Grants Council (project no. PolyU 11204919), and an internal research grant from the Hong Kong Polytechnic University (project 1.9B0V).

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th International Conference on Computational Semantics, pp. 13–22 (2013). https://www.aclweb.org/anthology/W13-0102/. Accessed 23 Oct 2020
2. Bengio, Y.: Discussion of the neural autoregressive distribution estimator. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, vol. 15, pp. 38–39 (2011). http://proceedings.mlr.press/v15/bengio11a/bengio11a.pdf. Accessed 22 Oct 2020
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
4. Cai, H., Chen, T., Zhang, W., Yu, Y., Wang, J.: Efficient architecture search by network transformation. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 2787–2794 (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16755. Accessed 11 Nov 2020
5. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with A-GEM. In: Proceedings of the 7th International Conference on Learning Representations (2019). https://openreview.net/forum?id=Hkf2_sC5FX. Accessed 1 Nov 2020
6. Chen, D., Mei, J., Wang, C., Feng, Y., Chen, C.: Online knowledge distillation with diverse peers. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, pp. 3430–3437 (2020). https://aaai.org/ojs/index.php/AAAI/article/view/5746. Accessed 1 Nov 2020
7. Chen, Q., Zhu, X., Ling, Z., Inkpen, D., Wei, S.: Neural natural language inference models enhanced with external knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2406–2417. Association for Computational Linguistics (2018). https://aclanthology.org/P18-1224/. Accessed 1 Jan 2021
8. Chen, T., Goodfellow, I.J., Shlens, J.: Net2net: Accelerating learning via knowledge transfer. In: Y. Bengio, Y. LeCun (eds.) Proceedings of the 4th International Conference on Learning Representations (2016). arxiv:1511.05641. Accessed 1 Jan 2021
9. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31th International Conference on Machine Learning, pp. 647–655 (2014)
10. Du, W., Black, A.W.: Data augmentation for neural online chats response selection. In: Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI, pp. 52–58 (2018). https://doi.org/10.18653/v1/w18-5708
11. Fan, W., Guo, Z., Bouguila, N., Hou, W.: Clustering-based online news topic detection and tracking through hierarchical bayesian nonparametric models. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2126–2130. ACM (2021). https://doi.org/10.1145/3404835.3462982
12. Feng, Y., Feng, J., Rao, Y.: Reward-modulated adversarial topic modeling. In: Proceedings of the 25th International Conference on Database Systems for Advanced Applications, vol. 12112, pp. 689–697 (2020). https://doi.org/10.1007/978-3-030-59410-7_47
13. Fu, Y., Feng, Y.: Natural answer generation with heterogeneous memory. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 185–195. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-1017
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems

vol 27, pp. 2672–2680 (2014). https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html. Accessed on 1 Sep 2020
15. Gupta, P., Chaudhary, Y., Buettner, F., Schütze, H.: Document informed neural autoregressive topic models with distributional prior. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pp. 6505–6512 (2019). https://doi.org/10.1609/aaai.v33i01.33016505
16. Gupta, P., Chaudhary, Y., Runkler, T.A., Schütze, H.: Neural topic modeling with continual lifelong learning. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 3907–3917 (2020). http://proceedings.mlr.press/v119/gupta20a.html. Accessed 10 Sep 2020
17. Han, X., Dai, Y., Gao, T., Lin, Y., Liu, Z., Li, P., Sun, M., Zhou, J.: Continual relation learning via episodic memory activation and reconsolidation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6429–6440 (2020). https://www.aclweb.org/anthology/2020.acl-main.573/. Accessed 1 Oct 2020
18. He, S., Liu, C., Liu, K., Zhao, J.: Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 199–208. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1019
19. Hida, R., Takeishi, N., Yairi, T., Hori, K.: Dynamic and static topic model for analyzing time-series document collections. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 516–520 (2018). https://www.aclweb.org/anthology/P18-2082/. Accessed 20 Sep 2020
20. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arxiv:1503.02531 (2015). Accessed 1 Sep 2020
21. Hoyle, A., Goel, P., Resnik, P.: Improving neural topic models using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 1752–1771 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.137
22. Hu, X., Wang, R., Zhou, D., Xiong, Y.: Neural topic modeling with cycle-consistent adversarial training. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 9018–9030 (2020). https://www.aclweb.org/anthology/2020.emnlp-main.725/. Accessed 25 Nov 2020
23. Huang, J., Peng, M., Li, P., Hu, Z., Xu, C.: Improving biterm topic model with word embeddings. World Wide Web **23**(6), 3099–3124 (2020). https://doi.org/10.1007/s11280-020-00823-w
24. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 448–456 (2015). http://proceedings.mlr.press/v37/ioffe15.html. Accessed 1 Oct 2020
25. Jiang, H., Zhou, R., Zhang, L., Wang, H., Zhang, Y.: A topic model based on poisson decomposition. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1489–1498. ACM (2017). https://doi.org/10.1145/3132847.3132942
26. Jiang, H., Zhou, R., Zhang, L., Wang, H., Zhang, Y.: Sentence level topic models for associated topics extraction. World Wide Web **22**(6), 2545–2560 (2019). https://doi.org/10.1007/s11280-018-0639-1
27. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.: Lightgbm: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 3146–3154 (2017). https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. Accessed on 20 Nov 2020
28. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. In: Proceedings of the 5th International Conference on Learning Representations (2017). https://openreview.net/forum?id=H1oyRlYgg. Accessed 1 Oct 2020
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (2015). arxiv: 1412.6980. Accessed 15 Sep 2020
30. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. In: Proceedings of the National Academy of Sciences, pp. 3521–3526 (2017)
31. Lauly, S., Zheng, Y., Allauzen, A., Larochelle, H.: Document neural autoregressive distribution estimation. Journal of Machine Learning Research **18**, 113:1–113:24 (2017). http://jmlr.org/papers/v18/16-017.html. Accessed 20 Sep 2020
32. Li, Z., Hoiem, D.: Learning without forgetting. In: Proceedings of the 14th European Conference on Computer Vision, pp. 614–629. Springer (2016)

33. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing **415**, 106–113 (2020). https://doi.org/10.1016/j.neucom.2020.07.048
34. Madotto, A., Wu, C., Fung, P.: Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1468–1478. Association for Computational Linguistics (2018). https://aclanthology.org/P18-1136/
35. Marsland, S., Shapiro, J., Nehmzow, U.: A self-organising network that grows when required. Neural Networks **15**, 1041–1058 (2002). https://www.sciencedirect.com/science/article/pii/S0893608000200078 3. Accessed 11 Nov 2020
36. Mccloskey, M.: Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation **24**, 109–165 (1989)
37. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering discrete latent topics with neural variational inference. In: Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 70, pp. 2410–2419 (2017). http://proceedings.mlr.press/v70/miao17a.html. Accessed 23 Sep 2020
38. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: Proceedings of the 33nd International Conference on Machine Learning, vol. 48, pp. 1727–1736 (2016). http://proceedings.mlr.press/v48/miao16.html. Accessed 20 Sep 2020
39. Mimno, D.M., Wallach, H.M., Talley, E.M., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011). https://www.aclweb.org/anthology/D11-1024/. Accessed 1 Oct 2020
40. li Ming, G., Song, H.: Adult neurogenesis in the mammalian brain: Significant answers and significant questions. Neuron **70**, 687–702 (2011). https://www.sciencedirect.com/science/article/pii/S0896627310003485. Accessed 15 Nov 2020
41. Mnih, A., Gregor, K.: Neural variational inference and learning in belief networks. In: Proceedings of the 31th International Conference on Machine Learning, JMLR Workshop and Conference Proceedings, pp. 1791–1799. JMLR.org (2014). http://proceedings.mlr.press/v32/mnih14.html. Accessed 10 Sep 2020
42. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning, pp. 807–814 (2010). https://icml.cc/Conferences/2010/papers/432.pdf. Accessed 11 Oct 2020
43. Nan, F., Ding, R., Nallapati, R., Xiang, B.: Topic modeling with wasserstein autoencoders. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, pp. 6345–6381 (2019). https://doi.org/10.18653/v1/p19-1640
44. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019). https://doi.org/10.1016/j.neunet.2019.01.012
45. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014). https://doi.org/10.3115/v1/d14-1162
46. Peters, M.E., Neumann, M., IV, R.L.L., Schwartz, R., Joshi, V., Singh, S., Smith, N.A.: Knowledge enhanced contextual word representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 43–54. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1005
47. Rebuffi, S., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the 30th Conference on Computer Vision and Pattern Recognition, pp. 5533–5542 (2017). https://doi.org/10.1109/CVPR.2017.587
48. Robins, A.V.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connect. Sci. **7**(2), 123–146 (1995). https://doi.org/10.1080/09540099550039318
49. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015). https://doi.org/10.1145/2684822.2685324
50. Shen, Y., Zeng, X., Jin, H.: A progressive model to enable continual learning for semantic slot filling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1279–1284 (2019). https://doi.org/10.18653/v1/D19-1126
51. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In: Proceedings of the 5th International Conference on Learning Representation (2017). https://openreview.net/forum?id=BybtVK9lg. Accessed 19 Sep 2020

52. Venkatesaramani, R., Downey, D., Malin, B.A., Vorobeychik, Y.: A semantic cover approach for topic modeling. In: Proceedings of the 8th Joint Conference on Lexical and Computational Semantics, pp. 92–102 (2019). https://doi.org/10.18653/v1/s19-1011
53. Wang, H., Xiong, W., Yu, M., Guo, X., Chang, S., Wang, W.Y.: Sentence embedding alignment for lifelong relation extraction. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 796–806 (2019). https://doi.org/10.18653/v1/n19-1086
54. Wang, R., Hu, X., Zhou, D., He, Y., Xiong, Y., Ye, C., Xu, H.: Neural topic modeling with bidirectional adversarial training. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 340–350 (2020). https://www.aclweb.org/anthology/2020.acl-main.32/. Accessed 19 Sep 2020
55. Wang, R., Zhou, D., He, Y.: ATM: adversarial-neural topic model. Information Processing and Management **56** (2019). https://doi.org/10.1016/j.ipm.2019.102098
56. Wang, R., Zhou, D., He, Y.: Open event extraction from online text using a generative adversarial network. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 282–291 (2019). https://doi.org/10.18653/v1/D19-1027
57. Wang, S., Chen, Z., Liu, B.: Mining aspect-specific opinion using a holistic lifelong topic model. In: Proceedings of the 25th International Conference on World Wide Web, pp. 167–176 (2016). https://doi.org/10.1145/2872427.2883086
58. Yang, P., Li, L., Luo, F., Liu, T., Sun, X.: Enhancing topic-to-essay generation with external commonsense knowledge. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, pp. 2002–2012. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1193
59. Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., Jiang, M.: A survey of knowledge-enhanced text generation. arxiv:2010.04389 (2020)
60. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning, pp. 3987–3995 (2017)
61. Zhang, H., Liu, Z., Xiong, C., Liu, Z.: Grounded conversation generation as guided traverses in commonsense knowledge graphs. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2031–2043 (2020). https://doi.org/10.18653/v1/2020.acl-main.184
62. Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X.: Commonsense knowledge aware conversation generation with graph attention. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 4623–4629. ijcai.org (2018) https://doi.org/10.24963/ijcai.2018/643