International Journal of Computer Vision (2022) 130:1920–1937
https://doi.org/10.1007/s11263-022-01624-6

Matching Visual Features to Hierarchical Semantic Topics for Image Paragraph Captioning

Dandan Guo[1] · Ruiying Lu[1] · Bo Chen[1] · Zequn Zeng[1] · Mingyuan Zhou[2]

Received: 7 April 2021 / Accepted: 27 April 2022 / Published online: 27 May 2022

Abstract

Observing a set of images and their corresponding paragraph-captions, a challenging task is to learn how to produce a semantically coherent paragraph to describe the visual content of an image. Inspired by recent successes in integrating semantic topics into this task, this paper develops a plug-and-play hierarchical-topic-guided image paragraph generation framework, which couples a visual extractor with a deep topic model to guide the learning of a language model. To capture the correlations between the image and text at multiple levels of abstraction and learn the semantic topics from images, we design a variational inference network to build the mapping from image features to textual captions. To guide the paragraph generation, the learned hierarchical topics and visual features are integrated into the language model, including Long Short-Term Memory and Transformer, and jointly optimized. Experiments on public datasets demonstrate that the proposed models, which are competitive with many state-of-the-art approaches in terms of standard evaluation metrics, can be used to both distill interpretable multi-layer semantic topics and generate diverse and coherent captions.

Keywords Image paragraph generation · Deep topic model · Language model · Image and text

1 Introduction

Describing visual content in a natural-language utterance is an emerging interdisciplinary problem, which lies at the intersection of computer vision (CV) and natural language processing (NLP) (Ordonez et al. 2016). As a sentence-level short image caption (Xu et al. 2015a; Vinyals et al. 2015; Anderson et al. 2018) has a limited descriptive capacity, Krause et al. (2017) introduce a paragraph-level captioning method that aims to generate a detailed and coherent paragraph for describing an image more finely. Recent advances in image paragraph generation focus on building different types of hierarchical recurrent neural network (HRNN), *e.g.*, LSTM (Hochreiter and Schmidhuber 1997), to generate the visual paragraphs. For HRNN, the high-level RNN takes the image features as the input and recursively produces a sequence of sentence-level vectors, which are often explained as the topic vectors; while the low-level RNN is subsequently adopted to decode each topic vector into an output sentence. By modeling each sentence and coupling the sentences into one paragraph, these hierarchical architectures often outperform existing flat models (Krause et al. 2017). To improve the performance and generate more diverse paragraphs, advanced methods, extending HRNN based on generative adversarial networks (GAN) (Goodfellow et al. 2014) or variational auto-encoders (VAE) (Kingma and Welling 2014), are proposed by Liang et al. (2017) and Chatterjee and Schwing (2018). Apart from adopting the output of the high-level RNN to represent the topics, Wang et al.

Communicated by Kwan-Yee Kenneth Wong.

Dandan Guo and Ruiying Lu have contributed equally to this work.

✉ Bo Chen
bchen@mail.xidian.edu.cn

Dandan Guo
gdd_xidian@126.com

Ruiying Lu
ruiyinglu_xidian@163.com

Zequn Zeng
zzq8341@gmail.com

Mingyuan Zhou
mingyuan.zhou@mccombs.utexas.edu

[1] National Laboratory of Radar Signal Processing, Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China

[2] McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA

(2019) introduce convolutional auto-encoding (CAE) on the region-level features of an image to learn the corresponding topics, which are further integrated into the HRNN-based paragraph generation framework.

In summary, the above image paragraph captioning methods typically refer to the output of the high-level RNN or CAE as the topics. Note that these topics are very different from the semantic topics represented by a set of semantically related words, explored in topic models (TMs). Designing topics in this way may cause these above models only attend to some visually salient image regions without grasping the image's main semantic topic. As discussed by Zhu et al. (2018), having an intuition about an image's high-level semantic topics is generally beneficial for selecting the most semantically-meaningful and topic-relevant image areas for describing an image. Recently, there are several attempts to utilize semantic topics learned from topic models, *e.g.*, Latent Dirichlet Allocation (LDA) (Blei et al. 2003), a commonly used shallow topic model, to generate a single-sentence caption (Fu et al. 2017; Zhu et al. 2018; Yu et al. 2018; Mao et al. 2018). Similar to them, Mao et al. (2018) extract textual topics of images with LDA, and generate topic-oriented multiple sentences to describe an image. Although having been proved effective, these above methods of utilizing semantic topics still have clear limitations. Typically, they only utilize shallow topic models to extract the single-layer semantic topics, which may have limited representation capacity. Another key limitation lies in adopting a two-stage manner to extract the semantic topics from images. Usually, they pre-train LDA from the captions of the training images and then train a downstream topic estimator to predict the topics with the image features as the input. However, this two-stage way does not consider the visual image features when learning the topic information and discard the uncertainty of the topic information brought by the probabilistic topic model, which is a desired property to capture the inherent ambiguity of paragraph generation from images.

This paper presents a flexible hierarchical-topic-guided image paragraph generation framework in an end-to-end manner, coupling a visual extractor with a multi-stochastic-layer deep topic model to guide the generation of a language model (LM). Specifically, a convolutional neural network (CNN) coupled with a region proposal network (Ren et al. 2015) is first utilized to detect a set of salient image regions as the visual extractor, a usual practice in image captioning systems. Motivated by the idea of using multi-layer features in Xu et al. (2015b) and Zhu et al. (2022), we aim to extract the multi-stochastic-layer topics and use them to guide the paragraph generation semantically, which has not been well exploited in existing methods for image paragraph captioning. To this end, we construct a deep topic model to match the image's visual features to its corresponding semantic topic information. Here we design a deep topic model built on the success of the Poisson gamma belief network (PGBN) (Zhou et al. 2016), which extracts interpretable multi-layer topics from text data and can be equivalently represented as deep LDA (Cong et al. 2017). A naive approach by introducing PGBN into the image paragraph caption task is adopting the two-stage manner similar to Mao et al. (2018), where we can pre-train PGBN, extract hierarchical topics from the training captions, and then build a downstream topic classifier with a deterministic network to approximate the deep topics from the image features. However, this inflexible way does not incorporate the visual image features into the learning process of the hierarchical topics and abandons the uncertainty in the probabilistic topic model (PGBN), resulting in unsatisfactory image paragraph captioning performance. To capture the correlations between the image and text at multiple levels of abstraction and learn the semantic topics from images, we here generalize PGBN into a novel visual-textual coupling model (VTCM). Generally, VTCM encapsulates region-level features into the hierarchical topics by a variational encoder and feeds the topic usage information to the decoder (PGBN) to generate descriptive captions. Different from existing image paragraph caption methods that compute topic vectors via deterministic neural networks (RNN or CAE) or learn the semantic topics using the shallow topic models (typically in a two-stage manner), our proposed VTCM can relate semantic topics and visual concepts and distill the multi-stochastic-layer topic information in an "end-to-end" manner.

To guide paragraph generation, both the visual features and mined hierarchical semantic topics from the VTCM are fed into either an LSTM or Transformer (Vaswani et al. 2017) based language generator. We refer to them as VTCM-LSTM and VTCM-Transformer. Following Wang et al. (2019), the LM in VTCM-LSTM capitalizes on both paragraph-level and sentence-level LSTMs. Inspired by the idea of selecting top-relevant regions (Anderson et al. 2018; Fan et al. 2020), the feedback of the paragraph-level LSTM is fed into the attention module together with the topic information to select critical image visual features. The sentence-level LSTM generates a sequence of words conditioning on the learned topics and attended image features. For Transformer-based image caption systems, while the original Transformer architecture can be directly adopted as the LM in our framework, the multi-modal nature of image captions requires specialized architectures different from those employed for the understanding of a single modality. Cornia et al. (2020) thus introduce a Meshed-Transformer with memory for image captioning, which learns a multi-level representation of the relationships between image regions via an augmented-memory encoder and uses mesh-like connectivity at the decoding stage to exploit both low- and high-level features. Our work aims to improve the Meshed-Transformer with the multi-stochastic-layer topic information, which is hierarchically coupled with the visual features extracted by the encoder and further interpolated into the feedback of each decoding layer to guide the caption generation. Absorbing the multi-layer semantic topics as additional guidance, both VTCM-LSTM and VTCM-Transformer produce a caption closely related to the given image and semantic topics. Unlike previous works that adopt GAN or VAE to enforce diversity in the generated captions (Liang et al. 2017; Chatterjee and Schwing 2018), our paragraph captioning systems can generate diverse captions for an image since we feed the multi-stochastic-layer latent topic representation of VTCM as the source of randomness to the language generator. Moving beyond existing methods that often use pre-trained semantic topics learned from LDA (Fu et al. 2017; Mao et al. 2018; Yu et al. 2018; Chen et al. 2019), our model allows jointly training the proposed VTCM (a deep topic model) and LM in an end-to-end manner. Since the semantic topics learned with VTCM are represented with a set of keywords, we can designate different topics as high-level guiding information, where the generated captions can be not only related to the image but also reflect what the user wants to emphasize. To the best of our knowledge, we are the first to distill the hierarchical semantic topics by capturing the correlations between the image and text at multiple levels of abstraction, and feed the topics into an LSTM-based or Transformer-based LM in an end-to-end manner to guide the paragraph generation. Due to the effectiveness and flexibility of our proposed plug-and-play system, one can also replace the language model with other architectures. Our main contributions include: 1) VTCM is proposed to extract and relate the hierarchical semantic topics with image features, where the distilled topics are integrated into both LSTM-based and Transformer-based LMs, guiding paragraph-level caption generation; 2) An end-to-end training is introduced to optimize the VTCM and LM jointly, beneficial for relating the visual and semantic concepts; 3) Extensive experiments are performed, with the quantitative and qualitative results showing the benefits of extracting multi-layer semantic topics for generating descriptive paragraphs.

2 Related Work

Below we review related work on image paragraph captioning, topic molding, and language modeling.

2.1 Image Paragraph Captioning

Image captioning aims to describe images with natural language, in which a popular research line is generating a single sentence to depict an image (Vinyals et al. 2015; Xu et al. 2015a), denoted as image sentence captioning. However, as a single-sentence description is often too short to capture all detailed information, image paragraph captioning has been proposed to describe an image by generating a paragraph consisting of multiple sentences. Besides, other research directions in image captioning have also gradually attracted attention. For example, visual storytelling (Tang et al. 2019) aims to generate narrative creations from ordered photo sequences. In addition, optical character recognition (OCR) based image captioning (Wang et al. 2021) aims to automatically describe an image with a sentence according to all the visual entities (both visual objects and scene text) in the image. While these problems are also challenging and attractive, they are beyond the scope of this work that is focused on image paragraph captioning. Regions-Hierarchical (Krause et al. 2017) designs HRNN to produce a generic paragraph for an image. To generate diverse and semantically coherent paragraphs, Liang et al. (2017) extend the HRNN by proposing an adversarial framework between structured paragraph generator and multi-level paragraph discriminators. Considering the difficulties associated with training GANs and deficiency of explicit coherence model, Chatterjee and Schwing (2018) augment HRNN with coherence vectors and a formulation of VAE (Kingma and Welling 2014). To encapsulate region-level features of an image into the topics, Wang et al. (2019) design a convolutional auto-encoding (CAE) module for topic modeling, where the extracted topics are further integrated into a two-level LSTM-based paragraph generator. Motivated by some models that utilize semantic topics to generate single-sentence captions (Fu et al. 2017; Zhu et al. 2018; Yu et al. 2018), Mao et al. (2018) pre-train the LDA (Blei et al. 2003) from the caption corpus of the training images at the first step, then train a topic classifier for semantic regularization and topic prediction based on the learned topics. In short, most of these existing paragraph captioning models either adopt deterministic networks (*e.g.*, high-level RNN or CAE) to construct a topic for each sentence within the whole paragraph or utilize a pre-trained LDA in a two-stage manner to learn the shallow semantic topics of images. Different from them, this work distills the multi-stochastic-layer semantic topics by capturing the correlations between the image and text at multiple levels of abstraction.

2.2 Topic Models and Language Models

Probabilistic topic models (PTMs), such as latent Dirichlet allocation (LDA) (Blei et al. 2003; Griffiths and Steyvers 2004) and Poisson factor analysis (PFA) (Zhou et al. 2012), often represent each document as a bag of words (BoW), capturing global semantic coherency into semantically meaningful topics. To explore the hierarchical semantic structures, PGBN (Zhou et al. 2016), a deep generalization of PFA that can also be viewed as a deep LDA (Cong et al. 2017), is proposed to extract interpretable hierarchical topics and capture the relationship between latent topics across multiple stochastic layers. Despite its effectiveness, PGBN, requiring the texts at training and testing stages, is not suitable for the image paragraph captioning task, where only images are given during the testing stage. To this end, we develop a variational encoder to match the visual features to the hierarchical topic information and generalize PGBN to build a novel visual-textual coupling model (VTCM), which can be jointly optimized with the paragraph generator. Although the idea of introducing a variational encoder is similar to neural topic models (NTMs) (Miao et al. 2016; Srivastava and Sutton 2017; Burkhardt and Kramer 2019; Zhang et al. 2018; Zhao et al. 2021), our VTCM captures the correlations between the image and text at multiple levels and is proposed to adapt to the image paragraph captioning task. To our knowledge, the works that connect deep topic modeling with visual features are still very limited. Different from Zhang et al. (2020) that propose to match visual features with a topic model, where the topics are fed into a GAN-based image generator, we focus on integrating the learned hierarchical topics from VTCM into the LMs to guide the paragraph generation, a task distinct from image generation.

Existing LMs are often built on either recurrent units, as used in RNNs (Cho et al. 2014; Hochreiter and Schmidhuber 1997), or purely the attention mechanism based modules, as used in Transformer and its various generalizations (Vaswani et al. 2017; Radford et al. 2018). RNN-based LMs have been successfully used in image paragraph captioning systems (discussed above), while single-sentence Transformer-based image captioning systems have started to attract attention (Huang et al. 2019; Li et al. 2019; Cornia et al. 2020), not to mention the research on image paragraph captioning. Note that we can flexibly select the LM for our plug-and-play system since we pay more attention to assimilating the multi-layer semantic topic information into the paragraph generator. We consider both the LSTM-based and Transformer-based LMs to investigate the effectiveness of integrating the deep semantic topics.

3 Proposed Models

Denoting *Img* as the given image, image paragraph captioning systems aim to generate a paragraph $P = \{S_1, ..., S_J\}$ consisting of J sentences, where sentence $S_j = \{w_{j,1}, ..., w_{j,T_j}\}$ consists of T_j words from a vocabulary of size V. We introduce a plug-and-play hierarchical-topic-guided image paragraph captioning system, with the overview of VTCM-LSTM depicted in Fig. 1. It contains three major components, including a visual extractor for extracting image features, the proposed VTCM for distilling multi-layer semantic topics of a given image, and the LM for interpreting the extracted image features and semantic topics into captions. Following a usual practice in image captioning systems, we implement the visual extractor by adopting a CNN coupled with a region proposal network (RPN) (Ren et al. 2015), as shown in Fig. 1a. The process is expressed as $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M\} = \mathrm{VE}(Img)$, where $\mathrm{VE}(\cdot)$ denotes the visual extractor, M the number of regions, $\boldsymbol{v}_i \in \mathbb{R}^D$ the i-th salient region of *Img* and D the dimension of visual features. To further compactly describe the content of the image, we subsequently aggregate these

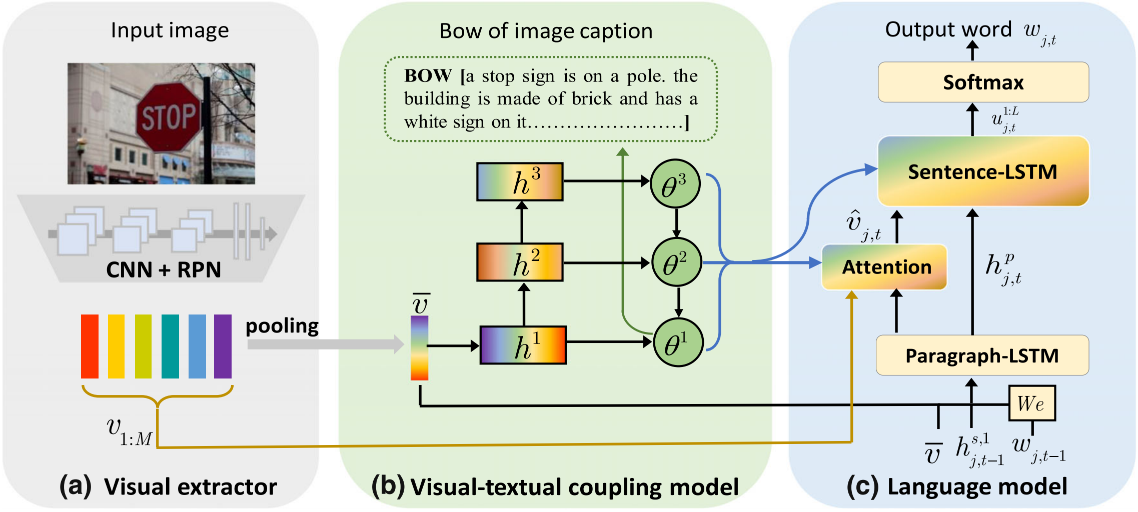


Fig. 1 Architecture of our proposed VTCM-LSTM. **a** The visual extractor, consisting of a CNN and an RPN, produces feature vectors $\boldsymbol{v}_{1:M}$ and average-pooled vector $\overline{\boldsymbol{v}}$. **b** The visual-textual coupling model, where the right part (from $\boldsymbol{\theta}^3$ to BoW of the paragraph caption) is the generative model with a three-hidden-layer (decoder) and left (from the average-pooled vector to the $\boldsymbol{h}^l$ and $\boldsymbol{\theta}^l$) is the variational encoder. **c** The LSTM-based LM, including paragraph-level LSTM, attention module and sentence-level LSTM, where $w_{j,t}$ is the t-th word in the j-th sentence of a paragraph and $\mathbf{W}_e$ is the word embedding matrix

M vectors into a single average-pooled vector $\overline{\boldsymbol{v}} = \frac{1}{M}\sum_i \boldsymbol{v}_i$. Below, we give more details about the other two components, *i.e.*, the VTCM and LM.

3.1 Visual-Textual Coupling Model

There are two mainstream ways to learn topics from an image. One is to encode the visual image features into a global vector with a high-level RNN, which is explained as the topic vectors and used to guide a low-level RNN. Note that these topics are very different from the semantic topics, which are usually represented in the form of semantically related words in a topic model. Another way is first applying LDA on the training captions and then training a downstream topic predictor over image features, where the semantic topics are assimilated into the LM for sentence or paragraph generation. Moving beyond them, we design an end-to-end variational deep topic model to capture the correlations between image features and descriptive text by distilling the semantic topics, jointly trained with the LM. The basic idea follows the philosophy that the generation from topics to descriptive captions via topic decoder and the topics extraction over image visual features via variational encoder can enforce the mined multi-layer topics to be related to the visual features.

Topic Decoder: As a multi-stochastic-layer deep generalization of LDA (Cong et al. 2017), PGBN (Zhou et al. 2016) is selected as the topic decoder. For the given *Img* in the training set, we summarize its ground-truth paragraph caption P into a BoW count vector $\boldsymbol{d} \in \mathbb{Z}_+^{V_c}$, where V_c is the size of the vocabulary excluding stop words, Z_+ denotes non-negative integers, and each element of $\boldsymbol{d}$ counts the number of times the corresponding word occurs in the paragraph. As shown in Fig. 1b, the generative process of PGBN with L-hidden-layer from top to bottom, is expressed as

$$\begin{aligned}
&\boldsymbol{\theta}^L \sim \mathrm{Gamma}\left(\boldsymbol{r}, \tau^{L+1}\right), \cdots, \\
&\quad \boldsymbol{\theta}^l \sim \mathrm{Gamma}\left(\boldsymbol{\Phi}^{l+1}\boldsymbol{\theta}^{l+1}, \tau^{l+1}\right), \cdots, \\
&\quad \boldsymbol{d} \sim \mathrm{Poisson}(\boldsymbol{\Phi}^1\boldsymbol{\theta}^1), \boldsymbol{\theta}^1 \sim \mathrm{Gamma}\left(\boldsymbol{\Phi}^2\boldsymbol{\theta}^2, \tau^2\right), \qquad (1)
\end{aligned}$$

where the shape parameters of the gamma distributed hidden units $\boldsymbol{\theta}^l \in \mathbb{R}_+^{K_l}$ are factorized into the product of connection weight matrix $\boldsymbol{\Phi}^{l+1} \in \mathbb{R}_+^{K_l \times K_{l+1}}$ and hidden units $\boldsymbol{\theta}^{l+1}$ of the next layer, K_l is the number of topics at layer l and $K_0 = V_c$. The vector $\boldsymbol{r} = \{r_1, ..., r_{K_L}\}$ at the top layer denotes the gamma shape parameter of $\boldsymbol{\theta}^L$; and $\{\tau^l\}_{l=1}^L$ are gamma scale parameters. We place a Dirichlet prior on each column of $\boldsymbol{\Phi}^l$ at each layer, denoted as $\boldsymbol{\phi}^l \sim \mathrm{Dir}(\eta^l, ..., \eta^l)$, where η^l is the prior of $\boldsymbol{\phi}^l$. The global semantics of image captions in training dataset are compressed into $\boldsymbol{\Phi}^{1:L}$, where $\boldsymbol{\Phi}^l \in \mathbb{R}_+^{K_{l-1} \times K_l}$ denotes K_l topics at layer l and each column of $\boldsymbol{\Phi}^l$ corresponds to a topic. To visualize the topic $\boldsymbol{\phi}_k^l$ at hidden layer l, we can map it to the V_c-dimensional observation space, expressed as $\left[\prod_{p=1}^{l-1}\boldsymbol{\Phi}^p\right]\boldsymbol{\phi}_k^l \in \mathbb{R}_+^{V_c}$, which is a distribution over all words in the vocabulary. $\boldsymbol{\theta}^{1:L}$ denote the hierarchical topic proportions of BoW count vector $\boldsymbol{d}$ over semantic topics $\boldsymbol{\Phi}^{1:L}$ and thus capture the semantic information of different levels about the given *Img*. Therefore, we can build a better paragraph generator by integrating hierarchical topic weight vectors into the language model.

Variational Topic Encoder: Under the hierarchical generative model of PGBN, conditioned on the captions of the training images, the inference task here is to find the global hierarchical topis $\boldsymbol{\Phi}^{1:L}$ (shared by all captions) and deep topic proportions $\boldsymbol{\theta}^{1:L}$ (specific for each caption). To infer $\boldsymbol{\Phi}^{1:L}$, we can adopt the topic-layer-adaptive stochastic gradient Riemannian Markov chain Monte Carlo (TLASGR-MCMC) developed by Cong et al. (2017) to provide a scalable distributional estimate. Given the topics $\boldsymbol{\Phi}^{1:L}$, inferring the topic proportions $\boldsymbol{\theta}^{1:L}$ from descriptive caption like the typical PGBN does is however not suitable for the image captioning task, where only images are given during the testing stage. To this end, a naive approach is adopting the two-stage manner. Namely, we can learn deep topics $\boldsymbol{\Phi}^{1:L}$ and topic proportions $\boldsymbol{\theta}^{1:L}$ using training captions and then learn a downstream topic predictor $f(\boldsymbol{\theta}^{1:L}|\overline{\boldsymbol{v}})$ to approximate $\boldsymbol{\theta}^{1:L}$ by taking visual features $\overline{\boldsymbol{v}}$ as the input. But, this two-stage way ignores the visual image features when learning the topic information and discards the uncertainty brought by the probabilistic topic model, a key property to capture the inherent ambiguity of paragraph generation from images. Therefore, motivated by the variational hetero-encoder in Zhang et al. (2020), we develop a variational topic encoder to match the visual features $\overline{\boldsymbol{v}}$ to the hierarchical topic weight vectors $\boldsymbol{\theta}^{1:L}$ and generalize PGBN into a novel visual-textual coupling model (VTCM). Specifically, we build a topic encoder as $\prod_{l=1}^L q(\boldsymbol{\theta}^l \mid \overline{\boldsymbol{v}})$, with

$$q(\boldsymbol{\theta}^l \mid \overline{\boldsymbol{v}}) = \mathrm{Weibull}(\boldsymbol{k}^l, \boldsymbol{\lambda}^l), \qquad (2)$$

where the Weibull distribution is used to approximate the gamma distributed conditional posterior, and its corresponding parameters $\boldsymbol{k}^l, \boldsymbol{\lambda}^l \in \mathbb{R}_+^{K_l}$ are deterministically nonlinearly transformed from the representation $\boldsymbol{h}^l$, which is mapped from the image features $\overline{\boldsymbol{v}}$, as described in "Appendix A.1" and shown in Fig. 1b. Using the reparameterization trick, we can sample the Weibull distributed topic weight vector $\boldsymbol{\theta}^l$ as

$$\boldsymbol{\theta}^l = \boldsymbol{\lambda}^l\left(-\ln(1-\boldsymbol{\epsilon}^l)\right)^{1/\boldsymbol{k}^l}, \ \boldsymbol{\epsilon}^l \sim \prod_{k=1}^{K_l}\mathrm{Uniform}(0,1). \qquad (3)$$

Benefiting from the variational framework, we can randomly draw $\boldsymbol{\theta}^l$ from the same latent space parameterized by $\boldsymbol{k}^l, \boldsymbol{\lambda}^l$, where different $\boldsymbol{\theta}^l$ can capture the inherent ambiguity (*i.e.*, diversity) for the given image but have the same semantic information. We denote $\boldsymbol{\Omega}_{\mathrm{TM}}$ as the set of encoder parameters, which can be updated via stochastic gradient descent (SGD) by maximizing a lower bound of the log marginal likelihood of caption $\boldsymbol{d}$ in (1), formulated as

$$L_{\mathrm{TM}} = \mathbb{E}_{q(\boldsymbol{\theta}^1 \mid \overline{\boldsymbol{v}})} \left[\ln p \left(\boldsymbol{d} \mid \boldsymbol{\Phi}^1 \boldsymbol{\theta}^1 \right) \right] - \sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{\theta}^l \mid \overline{\boldsymbol{v}})} \left[\ln q \left(\boldsymbol{\theta}^l \mid \overline{\boldsymbol{v}} \right) - \ln p \left(\boldsymbol{\theta}^l \mid \boldsymbol{\Phi}^{l+1} \boldsymbol{\theta}^{l+1} \right) \right]. \tag{4}$$

Optimizing the above lower bound will encourage the multi-stochastic-layer topic weight vectors $\boldsymbol{\theta}^{1:L}$ to capture holistic and representative information from the image and its corresponding caption. Serving as the bridge between two modalities, the hierarchical semantic topics can be further utilized to guide the caption generation of LMs. Note that we can flexibly select the LM for our plug-and-play system since we pay more attention to assimilating the multi-layer semantic topic weight vectors into the paragraph generator. Below we investigate how to integrate the topic information into not only LSTM-based but also Transformer-based LMs.

3.2 LSTM-based Language Generation Model

Inspired by Wang et al. (2019), who integrate the topics learned from CAE into a two-level LSTM-based paragraph generation framework with the attention mechanism in Anderson et al. (2018), we design a paragraph generator with a hierarchy constructed by a paragraph-level LSTM, a sentence-level LSTM, and an attention module, shown in Fig. 1c. The paragraph-level LSTM first encodes the semantic regions based on all previous words into the paragraph state. Then the attention module selects semantic regions with the guidance of the current paragraph state and semantic topics of the image. Finally, the sentence-level LSTM incorporates the topics, attended image features, and current paragraph state to facilitate word generation.

Paragraph-level LSTM: To generate $w_{j,t}$ as the t-th word of the j-th sentence in a paragraph caption, we set $\boldsymbol{x}^p_{j,t}$ as the input vector of the paragraph-level LSTM. By concatenating the previous output $\boldsymbol{h}^{s,1}_{j,t-1} \in \mathbb{R}^H$ of the sentence-level LSTM at layer 1, the image feature $\overline{\boldsymbol{v}}$, and previously generated word $w_{j,t-1}$, the $\boldsymbol{x}^p_{j,t}$ is formulated as $\boldsymbol{x}^p_{j,t} = \left[\boldsymbol{h}^{s,1}_{j,t-1}, \overline{\boldsymbol{v}}, \mathbf{W}_e w_{j,t-1}\right]$, where $[\cdot,\cdot]$ indicates concatenation, $\mathbf{W}_e \in \mathbb{R}^{E\times V}$ is a word embedding matrix, V the vocabulary size of LM, E the embedding size, and H the size of hidden state unit. This input provides paragraph-level LSTM the maximum contextual information, capturing both visual semantics of the image and long-range inter-sentence dependency within a paragraph caption (Anderson et al. 2018). Then the hidden state of the paragraph-level LSTM is computed as

$$\boldsymbol{h}^p_{j,t} = \mathrm{LSTM}_{\mathrm{para}}(\boldsymbol{x}^p_{j,t}, \boldsymbol{h}^p_{j,t-1}), \tag{5}$$

where $\boldsymbol{h}^p_{j,t} \in \mathbb{R}^H$ and $\boldsymbol{h}^p_{j,0} = \boldsymbol{h}^p_{j-1,T_{j-1}}$ are set to explore inter-sentence dependency.

Attention Module: Given the paragraph state $\boldsymbol{h}^p_{j,t}$ and the concatenation of multi-layer topic weight vectors $\boldsymbol{\theta}^{1:L}$, denoted as $[\boldsymbol{\theta}^{1:L}] \in \mathbb{R}_+^{\sum_{l=1}^L K_l}$, we build an attention module to select the most information-carrying regions of the visual features for predicting $w_{j,t}$, defined as

$$a^m_{j,t} = \boldsymbol{w}_{att} \tanh\left(\mathbf{W}_{va}\boldsymbol{v}_m + \mathbf{W}_{ha}\boldsymbol{h}^p_{j,t} + \mathbf{W}_{ta}[\boldsymbol{\theta}^{1:L}]\right),$$
$$\boldsymbol{p}_{j,t} = \mathrm{softmax}\left(\boldsymbol{a}_{j,t}\right), \tag{6}$$

where $a^m_{j,t}$ is the m-th element of $\boldsymbol{a}_{j,t} \in \mathbb{R}^M$, and $\boldsymbol{w}_{att} \in \mathbb{R}^{1\times A}$, $\mathbf{W}_{va} \in \mathbb{R}^{A\times D}$, $\mathbf{W}_{ha} \in \mathbb{R}^{A\times H}$, $\mathbf{W}_{ta} \in \mathbb{R}^{A\times(\sum_{l=1}^L K_l)}$ are the learned parameters and softmax is a function that turns the M-dimensional vector into a non-negative vector with M elements summed to 1. Defined in this way $\boldsymbol{p}_{j,t}$ is a probability vector over all regions in the image. The attended image feature is calculated with $\hat{\boldsymbol{v}}_{j,t} = \sum_{m=1}^M p^m_{j,t}\boldsymbol{v}_m$, providing a natural way to integrate the multi-layer semantic topics as auxiliary guidance when generating attention.

Sentence-level LSTM: The input vector $\boldsymbol{x}^s_{j,t}$ to the sentence-level LSTM at each time step consists of the output $\boldsymbol{h}^p_{j,t}$ of the paragraph-level LSTM, concatenated with the attended image feature $\hat{\boldsymbol{v}}_{j,t}$, stated as $\boldsymbol{x}^s_{j,t} = \left[\hat{\boldsymbol{v}}_{j,t}, \boldsymbol{h}^p_{j,t}\right]$. Specifically, sentence-level LSTM in turn produces a sequence of hidden states $\{\boldsymbol{h}^{s,l}_{j,1}, .., \boldsymbol{h}^{s,l}_{j,T_j}\} \in \mathbb{R}^H$ at layer l, one for each sentence in the paragraph, denoted as

$$\boldsymbol{h}^{s,l}_{j,t} = \begin{cases} \mathrm{LSTM}^l_{\mathrm{sent}}\left(\boldsymbol{h}^{s,l}_{j,t-1}, \boldsymbol{x}^s_{j,t}\right), & \text{if } l = 1, \\ \mathrm{LSTM}^l_{\mathrm{sent}}\left(\boldsymbol{h}^{s,l}_{j,t-1}, \boldsymbol{u}^{l-1}_{j,t}\right), & \text{if } L \geq l > 1, \end{cases} \tag{7}$$

where $\boldsymbol{u}^l_{j,t}$ is the coupling vector combining the topic weight vectors $\boldsymbol{\theta}^l$ and hidden output of the sentence-level LSTM $\boldsymbol{h}^{s,l}_{j,t}$ at each time step t. Following Guo et al. (2020), we realize $\boldsymbol{u}^l_{j,t} = g^l\left(\boldsymbol{h}^{s,l}_{j,t}, \boldsymbol{\theta}^l\right)$ with a gating unit similar to the gated recurrent unit (Cho et al. 2014), described in "Appendix A.2". The probability over words in the dictionary can be predicted by taking a linear projection and a softmax operation over the concatenation of $\boldsymbol{u}^l_{j,t}$ across all layers. This method enhances the representation power and, with skip connections from all hidden layers to the output (Graves et al. 2013), mitigates the vanishing gradient problem. We denote the parameters of the LSTM-based LM as $\boldsymbol{\Omega}_{\mathrm{LSTM}}$.

3.3 Transformer-based Language Generation Model

To demonstrate the proposed plug-and-play system, we also explore how to integrate the multi-layer topic information into existing Transformer-based LMs, due to their representation power and computational efficiency coming from pure attention mechanisms. Typically, attention operates on a set of queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$, defined as $\text{Attention}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(Q\boldsymbol{K}^T/\sqrt{d}\right)\boldsymbol{V}$, where $\boldsymbol{Q}$ is a matrix of n_q query vectors, both $\boldsymbol{K}$ and $\boldsymbol{V}$ contain n_k keys and values, all with the same dimensionality, and d is a scaling factor. Cornia et al. (2020) design a novel Transformer-based architecture to improve the image encoder and language decoder, and prove its effectiveness on sentence-level image captions. However, they neither consider the paragraph-level image captioning task nor the semantic topics underlying the image. On the basis of this Transformer-based architecture, we here devise a semantic topic-guided Transformer model, which is conceptually divided into an encoder and a decoder module, shown in Fig. 2. Following Cornia et al. (2020), the encoder processes region-level image features and devises the relationships between them. The decoder not only reads from the output of each encoding layer like that of Cornia et al. (2020) but also assimilates the hierarchical topic information to generate the paragraph caption word by word.

Memory-augmented Encoder: Denoting the aforementioned set of features $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M\}$ as $\mathbf{X}$ for clarity, we adopt the memory-augmented attention operator to encode image regions and their relationships, defined as

$$\mathcal{M}_{\text{mem}}\,(\boldsymbol{X}) = \text{ Attention }\left(W_q\boldsymbol{X}, \boldsymbol{K}, \boldsymbol{V}\right),$$
$$\boldsymbol{K} = [W_k\boldsymbol{X}, \boldsymbol{M}_k]\,, \quad \boldsymbol{V} = [W_v\boldsymbol{X}, \boldsymbol{M}_v]\,, \tag{8}$$

where W_q, W_k, W_v are matrices of learnable weights, a usual practice in the original Transformer, and $\boldsymbol{M}_k$ and $\boldsymbol{M}_v$ are additional keys and values implemented as plain learnable memory matrices. Following the implementation of Vaswani et al. (2017), the memory-augmented attention can be applied in a multi-head fashion, whose output can be fed into a feed-forward layer, denoted as $\mathcal{F}(\cdot)$. Both the attention and feed-forward layers are encapsulated within a residual connection and a layer norm operation, denoted as AddNorm$(\cdot)$. For the encoder with L layers, its l-th encoding layer is therefore defined as

$$\tilde{\boldsymbol{X}}^l = \text{AddNorm }\left(\mathcal{F}(\boldsymbol{Z}^l)\right),$$
$$\boldsymbol{Z}^l = \text{AddNorm }\left(\mathcal{M}_{\text{mem}}(\tilde{\boldsymbol{X}}^{l-1})\right), \tag{9}$$

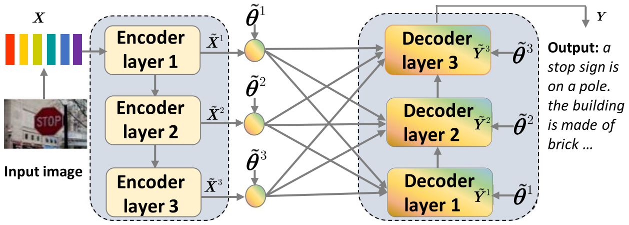


Fig. 2 The overview of VTCM-Transformer, where VTCM is the same topic model used in VTCM-LSTM and omitted here

where $\tilde{\boldsymbol{X}}^l \in R^d$ and $\tilde{\boldsymbol{X}}^0 = \boldsymbol{X}$. A stack of L encoding layers will produce a multilevel output $\tilde{\mathcal{X}} = \left(\tilde{\boldsymbol{X}}^1, \ldots, \tilde{\boldsymbol{X}}^L\right)$.

Topic-guided Meshed Decoder: Given the region encodings $\tilde{\mathcal{X}}$ and topic information $\boldsymbol{\theta}^{1:L}$, our decoder aims to generate the paragraph caption, denoted as $\boldsymbol{Y} = \{y_1, \ldots, y_I\}$ consisting of I words for clarity. Inspired by Cornia et al. (2020), we construct the topic-guided Meshed Attention operator to connect $\boldsymbol{Y}$ to all elements in $\tilde{\mathcal{X}}$ and $\boldsymbol{\theta}^{1:L}$ hierarchically through gated cross-attentions, formulated as

$$\mathcal{M}_{\text{mesh}}\,(\tilde{\mathcal{X}}, \tilde{\boldsymbol{\theta}}^{1:L}, \boldsymbol{Y}) = \sum_{l=1}^{L} \boldsymbol{\alpha}_l \odot \mathcal{C}\left(\tilde{\boldsymbol{X}}^l + \tilde{\boldsymbol{\theta}}^l, \boldsymbol{Y}\right), \tag{10}$$

where we combine the topic information $\tilde{\boldsymbol{\theta}}^l$ and the hidden output of the memory-augmented encoder $\tilde{\boldsymbol{X}}^l$ at each layer l although other choices are also available, $\tilde{\boldsymbol{\theta}}^l \in R^d$ is projected from $\boldsymbol{\theta}^l \in R^{K_l}$ into the encoder embedding space, and $\mathcal{C}(\cdot, \cdot)$ stands for the cross-attention, computed using queries from the decoder and keys and values from the encoder and topic information:

$$\mathcal{C}\left(\tilde{\boldsymbol{X}}^l + \tilde{\boldsymbol{\theta}}^l, \boldsymbol{Y}\right)$$
$$= \text{ Attention }\left(W_q\boldsymbol{Y}, W_k(\tilde{\boldsymbol{X}}^l + \tilde{\boldsymbol{\theta}}^l), W_v(\tilde{\boldsymbol{X}}^l + \tilde{\boldsymbol{\theta}}^l)\right). \tag{11}$$

By computing $\boldsymbol{\alpha}_l = \text{sigmoid}\left(W_l\left[\boldsymbol{Y}, \mathcal{C}\left(\tilde{\boldsymbol{X}}^l + \tilde{\boldsymbol{\theta}}^l, \boldsymbol{Y}\right)\right]\right)$, we can measure the relevance between cross-attention results, where W_l is the learned weight matrix. Similar to the encoding layer, the final structure of each decoding layer is written as

$$\tilde{\boldsymbol{Y}}^l = \text{AddNorm}(\mathcal{F}(\boldsymbol{Z}^l + \tilde{\boldsymbol{\theta}}^l)),$$
$$\boldsymbol{Z}^l = \text{AddNorm }\left(\mathcal{M}_{\text{mesh}}\left(\tilde{\mathcal{X}}, \tilde{\boldsymbol{\theta}}^{1:L}, \text{AddNorm}\left(\mathcal{S}_{\text{m}}(\tilde{\boldsymbol{Y}}^{l-1})\right)\right)\right), \tag{12}$$

where $\mathcal{S}_{\text{m}}$ is a masked self-attention used in the original Transformer (Vaswani et al. 2017), due to the prediction of a word should only depend on previously predicted words, and $\tilde{\boldsymbol{Y}}^0 = \boldsymbol{Y}$. After taking a linear projection and a softmax operation over $\tilde{\boldsymbol{Y}}^L$, our decoder finally predicts the probability over words in the vocabulary. Similar to the LSTM-based

LM, Transformer-based LM is also guided by the multi-layer semantic topics and attended image features when generating the caption, whose parameters are represented as $\boldsymbol{\Omega}_{\text{Trans}}$.

3.4 Joint Learning

Under the deep topic model described in Sect. 3.1 and LSTM-based LM in Sect. 3.2, the joint likelihood of the target ground truth paragraph P of Img and its corresponding BoW count vector $\boldsymbol{d}$ is defined as

$$p\left(P, \boldsymbol{d} \mid \boldsymbol{\Phi}^{1:L}, \boldsymbol{v}_{1:M}\right) = \int p\left(\boldsymbol{d} \mid \boldsymbol{\Phi}^{1}\boldsymbol{\theta}^{1}\right)\left[\prod_{l=1}^{L} p\left(\boldsymbol{\theta}^{l} \mid \boldsymbol{\Phi}^{l+1}\boldsymbol{\theta}^{l+1}\right)\right] \prod_{j=1}^{J}\prod_{t=1}^{T_j} p\left(w_{j,t} \mid w_{j,<t}, \boldsymbol{v}_{1:M}, \boldsymbol{\theta}^{1:L}\right) d\boldsymbol{\theta}^{1:L}, \tag{13}$$

which is similar as the likelihood of the topic-guided Transformer-based captioning system, described in "Appendix A.3". As discussed in Sect. 3.1, we introduce a variational topic encoder to learn the multi-layer topic weight vectors $\boldsymbol{\theta}^{1:L}$ in (2) with the image features as the input. Thus, a lower bound of the log of (13) can be constructed as

$$\begin{aligned} L_{\text{all}} &= \mathbb{E}_{q(\boldsymbol{\theta}^{1} \mid \overline{\boldsymbol{v}})}\left[\ln p\left(\boldsymbol{d} \mid \boldsymbol{\Phi}^{1}\boldsymbol{\theta}^{1}\right)\right] \\ &- \sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{\theta}^{l} \mid \overline{\boldsymbol{v}})}\left[\ln q\left(\boldsymbol{\theta}^{l} \mid \overline{\boldsymbol{v}}\right) - \ln p\left(\boldsymbol{\theta}^{l} \mid \boldsymbol{\Phi}^{l+1}\boldsymbol{\theta}^{l+1}\right)\right] \\ &+ \sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{\theta}^{l} \mid \overline{\boldsymbol{v}})}\left[\sum_{j=1}^{J}\sum_{t=1}^{T_j} \ln p\left(w_{j,t} \mid w_{j,<t}, \boldsymbol{v}_{1:M}, \boldsymbol{\theta}^{1:L}\right)\right], \end{aligned} \tag{14}$$

which unites the first two terms primarily responsible for training the topic model component, and the last term for training the LM component. The parameters $\boldsymbol{\Omega}_{\text{TM}}$ of the variational topic encoder and the parameters $\boldsymbol{\Omega}_{\text{LSTM}}$ of LSTM-based LM can be jointly updated by maximizing L_{all}. Besides, the global parameters $\boldsymbol{\Phi}^{1:L}$ of the topic decoder can be sampled with TLASGR-MCMC in Cong et al. (2017), described in "Appendix A.4". The training strategy is outlined in Algorithm 1.

To sum up, as shown in Fig. 1, the proposed framework couples the topic model (VTCM) with a visual extractor, which takes the visual features of the given image as input and maps the hierarchical topic weight vectors. The learned topic vectors in different layers are then used to reconstruct the BoW vector of the given image paragraph caption and as the additional features for the LSTM (or Transformer)-based LM to generate the paragraph. Moreover, our proposed model introduces randomness into the topic weight vector, which captures the uncertainty about what is depicted in an image and hence encourages the diversity of generation.

Algorithm 1 Inference for our proposed VTCM-LSTM.

```
Set mini-batch size $N$, the number of layer $L$ and the width of layer $K_l$;
Initialize topic encoder parameters $\boldsymbol{\Omega}_{\text{TM}}$ and LSTM-based parameters $\boldsymbol{\Omega}_{\text{LSTM}}$ and topic decoder parameters $\boldsymbol{\Phi}^{1:L}$.
for $iter = 1, 2, \cdots$ do
    Randomly select a mini-batch of $N$ images and their paragraph captions to form a subset $\{Img_n, P_n, \boldsymbol{d}_n\}_{n=1}^{N}$.
    Compute the image features with visual extractor;
    Draw random noise $\{\boldsymbol{\epsilon}_n^l\}_{n=1,l=1}^{N,L}$ from uniform distribution and sample latent states $\{\boldsymbol{\theta}_n^l\}_{n=1,l=1}^{N,L}$ from (3) via $\boldsymbol{\Omega}_{\text{TM}}$, which are fed into the LSTM with (6) and (7);
    Compute $\nabla_{\boldsymbol{\Omega}_{\text{TM}}} L_{\text{all}}$ and $\nabla_{\boldsymbol{\Omega}_{\text{LSTM}}} L_{all}$ according to 3.4, and update $\boldsymbol{\Omega}_{\text{TM}}$ and $\boldsymbol{\Omega}_{\text{LSTM}}$;
    Update $\boldsymbol{\Phi}^{1:L}$ with $\{\boldsymbol{\theta}_n^l\}_{n=1,l=1}^{N,L}$, described in "Appendix A.4";
end for
```

4 Experiments

4.1 Dataset and Implementation Details

We conduct experiments on the public Stanford image-paragraph dataset (Krause et al. 2017), where 14,575 image-paragraph pairs are used for training, 2,487 for validation, and 2,489 for testing. Following the standard evaluation protocol, we use the full set of captioning metrics: METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam et al. 2015), and BLEU (Papineni et al. 2002). Different from the BLEU scores primarily measuring the n-gram precision, METEOR and CIDEr are known to provide more robust evaluations of language generation algorithms (Vedantam et al. 2015). In our experiments, the hyper-parameters and model checkpoints are chosen by optimizing the performance based on the average of METEOR and CIDEr scores on the validation set.

Following the publicly available implementation of Anderson et al. (2018) and Wang et al. (2019), we use Faster R-CNN (Ren et al. 2015) with VGG16 network (Simonyan and Zisserman 2015) as the visual extractor, which is pre-trained over Visual Genome (Krishna et al. 2017). The top $M = 50$ detected regions are selected to represent image features. The size of each image feature vector is 4096, which is embedded into the 1024-dimensional vector before being fed into our topic model. For our LMs, we tokenize words and sentences using Stanford CoreNLP (Manning et al. 2014), lowercase all words, and filter out words that occur less than 1 time. We set the maximum number of sentences in each paragraph as 6 and the maximum length of each sentence as 30 (padded where necessary) for VTCM-LSTM. For our topic model, all the words from the training dataset, excluding the stopwords and the top 0.1% most frequent words, are used to obtain a BoW caption for the corresponding image. The hidden sizes of paragraph-LSTM, sentence-LSTM, and attention module are all set to 512. For our Transformer-based LM, we set the dimensionality d of each layer as 512, the number of heads as 8, and the number of memory vectors as 40. Both our Transformer-based and LSTM-based LMs are a three-layer model, same with the topic model with the topic number of $[K_1; K_2; K_3] = [80; 50; 30]$. Besides, we directly set hyper-parameters in VTCM as $\{\eta^l = 0.1, \tau^l = 1, \boldsymbol{r} = 1\}$. We use the Adam optimizer (Kingma and Ba 2015) with a learning rate of $5e - 4$ for VTCM-LSTM and 1 for VTCM-Transformer. The gradients of both VTCM-LSTM and VTCM-Transformer are clipped if the norm of the parameter vector exceeds 0.1. The dropout rate is set to 0.5 for VTCM-LSTM and 0.9 for VTCM-Transformer, and adopted in both the input and output layers to avoid overfitting. During inference, we adopt the penalty on trigram repetition proposed by Melas-Kyriazi et al. (2018) and set the penalty hyperparameter as 2. We also provide additional experimental results on the radiology report generation task in the "Appendix A.5".

4.2 Baselines

For a fair comparison, we consider the following baselines: (1) Image-Flat (Karpathy and Fei-Fei 2015), directly decoding a paragraph word-by-word via a single LSTM; (2) Flat-repetition-penalty (Melas-Kyriazi et al. 2018), training the non-hierarchical LSTM-based LM with an integrated penalty on trigram repetition to improve the diversity in image paragraph captioning; (3) Regions-Hierarchical (Krause et al. 2017), using a hierarchical LSTM to generate a paragraph, sentence by sentence; (4) RTT-GAN (Liang et al. 2017), training the Regions-Hierarchical in a GAN setting, coupled with an attention mechanism; (5) TOMS (Mao et al. 2018), generating multi-sentences under the topic guidence, which trains a downstream topic classifier to predict the topics mined by the LDA; (6) Diverse-VAE (Chatterjee and Schwing 2018), leveraging coherence vectors and global topic vectors to generate paragraph, under a VAE framework; (7) IMAP (Xu et al. 2020), proposing an interactive key-value memory-augmented attention into the hierarchical LSTM; (8) LSTM-ATT, which refers the outputs of paragraph-level LSTM as the topic vectors and adopts the attention mechanism in Anderson et al. (2018), as a degraded version of the proposed VTCM-LSTM; (9) $\mathcal{M}^2$-Transformer (Cornia et al. 2020), a novel Transformer-based architecture for single-sentence image captioning and a degraded version of VTCM-Transformer; 10) CAE-LSTM (Wang et al. 2019), which adopts the CAE to extract topics and integrates them into the two-level LSTM-based paragraph generator; 11) Splitting to Tree Decoder (S2TD) (Shi et al. 2021), which models the paragraph decoding process as a top-down binary tree expansion and consists of a split module, a score module, and a word-level LSTM; (12) Retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning (RAMP) (Xu et al. 2021), which makes full use of the R-best retrieved candidate captions and adopts the hierarchical LSTM as the paragraph generator.

4.3 Quantitative Evaluation

Main Results: The results of different models on the Stanford dataset are shown in Table 1, where we only report the results of different models trained with cross-entropy rather than self-critical sequence training to eliminate the influence of different training strategies. As it can be observed, our proposed VTCM-LSTM surpasses all the other LSTM-based captioning systems in terms of BLEU-4, BLEU-3, and CIDEr, while being competitive on BLEU-1, BLEU-2 and METEOR with the best performer. Moreover, on all metrics, our proposed VTCM-LSTM and VTCM-Transformer improve their corresponding baselines, *i.e.*, LSTM-ATT and $\mathcal{M}^2$-Transformer, respectively. These results demonstrate the effectiveness of integrating the semantic topics mined from VTCM into language generation in terms of topical semantics and descriptive completeness. Moreover, VTCM-Transformer leads to a performance boost over VTCM-LSTM on almost all the metrics, indicating the advantage of the memory-augmented operator and meshed cross-attention operator with a Transformer-like layer. We also replace the $\mathcal{M}^2$-Transformer with the original Transformer pretrained on a diverse set of unlabeled text, which however produces poor performance, suggesting the importance of designing specialized architectures for multi-model image captioning. Of particular note is the large improvement under both VTCM-LSTM and VTCM-Transformer on CIDEr, which is proposed specifically for image descriptions evaluation and measures the n-gram accuracy by term-frequency inverse-document-frequency (TF-IDF). Interestingly, by bridging the visual features to the textual descriptions, our proposed VTCM is suited for extracting paragraph-level word concurrence patterns into latent topics, which capture the main aspects of the scene and image descriptions. The assimilation of topic information into language models thus leads to a large improvement in CIDEr, correlated well with human judgment. However, it is often not the case in other image captioning systems unless the CIDEr score is treated as the reward and directly optimized with policy-gradient based reinforcement learning techniques to finetune the model (Wang et al. 2019; Cornia et al. 2020; Melas-Kyriazi et al. 2018; Xu et al. 2020).

Notably, the CAE-LSTM of Wang et al. (2019), the S2TD of Shi et al. (2021), and the RAMP of Xu et al. (2021) additionally adopt self-critical (SC) training after the pre-training with cross-entropy (CE). For a fair comparison, we also treat the CIDEr as the reward and introduce the self-critical into the pre-trained VTCM-LSTM, where the results are reported in Table 2. Our proposed model outperforms all the baselines on METEOR and CIDEr and achieves a comparable perfor-

Table 1 Main results for generating paragraphs Krause et al. (2017)

Method	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Image-Flat (Karpathy and Fei-Fei 2015)	12.82	11.06	34.04	19.95	12.20	7.71
Flat-repetition-penalty (Melas-Kyriazi et al. 2018)	15.17	22.68	35.68	22.40	14.04	8.70
TOMS (Mao et al. 2018)	18.6	20.8	43.1	25.8	14.3	8.4
Regions-Hierarchical (Krause et al. 2017)	15.95	13.52	41.90	24.11	14.23	8.69
RTT-GAN (Liang et al. 2017)	17.12	16.87	41.99	24.86	14.89	9.03
Diverse-VAE (Chatterjee and Schwing 2018)	**18.62**	20.93	42.38	25.52	15.15	9.43
IMAP (Xu et al. 2020)	16.56	20.76	42.38	25.87	15.51	9.42
S2TD(Shi et al. 2021)	17.00	21.92	**44.59**	**26.06**	14.93	8.35
LSTM-ATT	17.40	20.11	40.8	24.75	14.81	8.95
Our VTCM-LSTM	17.52	**22.82**	42.80	25.50	**15.69**	**9.63**
$\mathcal{M}^2$ Transformer (Cornia et al. 2020)	15.4	16.1	37.5	22.3	13.7	8.4
Our VTCM-Transformer	16.88	**26.15**	40.93	25.51	**15.94**	**9.96**
Human	19.22	28.55	42.88	25.68	15.55	9.66

The best results from LSTM-based LM and Transformer-based LM are marked in bold, respectively
Our models are compared with competing baselines along with six language metrics. The human performance is included for providing a better understanding of all metrics following

Table 2 Comparison of the proposed VTCM-LSTM and baselines trained with different strategies

Method	METEOR	CIDEr	BLEU-4
Our VTCM-LSTM (CE)	17.52	22.82	9.63
CAE-LSTM (CE+SC) (Wang et al. 2019)	18.82	25.15	9.67
S2TD(CE+SC)(Shi et al. 2021)	17.64	24.33	10.17
RAMP (CE+SC)(Xu et al. 2021)	17.49	23.22	**10.48**
Our VTCM-LSTM (CE+SC)	**18.95**	**25.50**	9.88

The best results from LSTM-based LM and Transformer-based LM are marked in bold, respectively

mance with S2TD and RAMP on BLEU-4. It indicates the effectiveness of assimilating the hierarchical semantic topic from VTCM into the paragraph generator.

Ablation Study: Firstly, we investigate the impact of topic layers on captioning performance. As it can be seen in Table 3, our proposed VTCM-LSTM and VTCM-Transformer can produce the desired improvement as its number of layers increases, showcasing the benefits of extracting multi-layer semantic topics for generating descriptive paragraphs. Secondly, to compare our proposed VTCM with other topic models, we adopt PGBN and two representative NTMs as variants to adapt the image paragraph generation task, where LDA with product of experts (ProdLDA) of Srivastava and Sutton 2017 presents the effective VAE-based inference algorithm for LDA and uses logistic normal distribution for the Dirichlet prior and Dirichlet VAE (DVAE) of Burkhardt and Kramer 2019 introduces a novel method based on rejection sampling variational inference. Specifically, for PTM, we first train PGBN on the training captions and learn a downstream topic estimator over image features to approximate the topic proportions $\boldsymbol{\theta}^{1:L}$ in a two-stage manner; for NTMs, including ProdLDA and DVAE, we replace the BoW representation with the visual features as the input into encoder and optimize them with the paragraph generator jointly, where both ProdLDA and DVAE can only extract shallow semantic topics. Although being effective, PGBN-LSTM and PGBN-Transformer are still inferior to their respective opponents using VTCM, which suggests the benefit of introducing the variational topic encoder. The reason behind this might be that the variational topic encoder learns the hierarchical semantic topics by capturing the correlations between image features and descriptive text jointly. Even though the variants of ProdLDA and DVAE can achieve comparable performance with our proposed VTCM-LSTM and VTCM-Transformer at layer 1, they can not capture the multi-stochastic-layer semantic topics like our VTCM, limiting their ability to generate more coherent paragraphs. Last,to evaluate the effectiveness of our way for integrating the topic information into the LMs, we provide two simple variants of our proposed models, $i.e.$, Topic+LSTM and Topic+Transformer where the topic information is directly concatenated to the output of ahead of the softmax at each time step, based on our adopted hierarchical LSTM and $\mathcal{M}^2$-Transformer. The proposed LSTM-based and Transformer-based models

Table 3 Ablation study on Stanford dataset

Method	M	C	B1	B2	B3	B4
ProdLDA-LSTM	16.22	18.52	42.32	24.86	15.07	9.07
DVAE-LSTM	15.77	18.35	41.63	24.47	14.44	8.96
PGBN-LSTM L=1	16.15	18.69	42.02	24.83	15.11	9.04
PGBN-LSTM L=2	16.37	19.28	42.36	24.80	15.22	9.23
PGBN-LSTM L=3	17.16	20.49	42.64	25.17	15.23	9.27
VTCM-LSTM L=1	16.50	19.10	42.39	25.42	15.41	9.33
VTCM-LSTM L=2	16.66	19.98	42.45	25.39	15.46	9.34
VTCM-LSTM L=3	**17.52**	**22.82**	**42.80**	**25.50**	**15.69**	**9.63**
ProdLDA-Transformer	15.56	21.45	39.76	22.36	14.18	8.54
DVAE-Transformer	15.21	21.12	39.24	21.87	13.96	8.32
PGBN-Transformer L=1	15.41	21.35	39.55	22.40	14.17	8.51
PGBN-Transformer L=2	16.18	23.54	40.11	23.25	14.72	9.08
PGBN-Transformer L=3	16.22	24.83	40.61	25.33	15.51	9.96
VTCM-Transformer L=1	15.87	22.71	39.61	22.92	14.21	8.65
VTCM-Transformer L=2	16.31	23.86	40.17	23.74	15.01	9.16
VTCM-Transformer L=3	**16.88**	**26.15**	**40.93**	**25.51**	**15.94**	**9.96**
Topic+LSTM L=3	15.47	18.02	41.80	24.61	14.74	9.10
Topic+Transformer L=3	15.66	23.45	38.77	23.14	14.51	8.87

The best results from LSTM-based LM and Transformer-based LM are marked in bold, respectively
Here, M, C and BN are short for the METEOR, CIDEr and BLEU-N, respectively

both outperform their corresponding base variants, which clearly indicates the usefulness of our proposed ways of incorporating the multi-layer semantic topics into the language decoding process.

4.4 Qualitative Evaluation

Generated Captions: To qualitatively show the effectiveness of our proposed methods, we show descriptions of different images generated by different methods in Fig. 3. As we can see, all of these models can produce paragraphs related to the given images, while our proposed VTCM-LSTM and VTCM-Transformer can generate more coherent and accurate paragraphs by learning to distill the semantic topics from an image via the VTCM module to guide paragraph generation. Therefore, instead of only attending to some visually salient image regions, the generated descriptions of our models are also highly related to the given images in terms of their semantic meanings but not necessarily the words same as the original caption. Taking the first row as the example, the proposed VTCM-Transformer can generate coherent and meaningful paragraphs to describe the image, while capturing meta-concepts like “steam train” and “driving on” based on the scenes including “train” and “smoke”. Notably, these concepts are even not described in the ground truth but are very relative to the whole image. However, without the high-level semantic information, these baselines tend to only describe all the salient visual objects in the image, ignoring the “main plot” underlying the images, such as the “the train is in motion” in the first row and “food” in the last row in Fig. 3. These observations suggest that the proposed VTCM has successfully captured hierarchical semantic topics by matching visual features to descriptive texts with a similar VAE structure, and our proposed ways of assimilating the topic information into LSTM or Transformer can successfully guide the paragraph generation.

Learned Topics: One of the benefits of introducing the hierarchical semantic topics learned from VTCM is the enhancement of model interpretability. To examine whether the topic model can learn the desired topics from the input image, in Fig. 4, we visualize the learned hierarchical topics with our topic model given the input images from the test set, where each topic in different layers has a list of representative words with decreasing ranks. It is clear that the extracted multi-layer topics are highly correlated with the chosen image and its corresponding text. In other words, the topic model we use successfully capture the semantically related topics given the image features. However, the interpretability of semantic topics is not possessed by the topics in most existing image paragraph captioning models, which are computed by RNN or CAE. Besides, we can see that the topics become more and more specific when moving from the top layer to the bottom layer. Furthermore, we note that the same scene from different topics, such as the word “brown,” is described from multiple perspectives, making it possible to describe the input image from different topic perspectives.

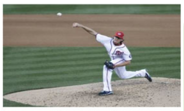
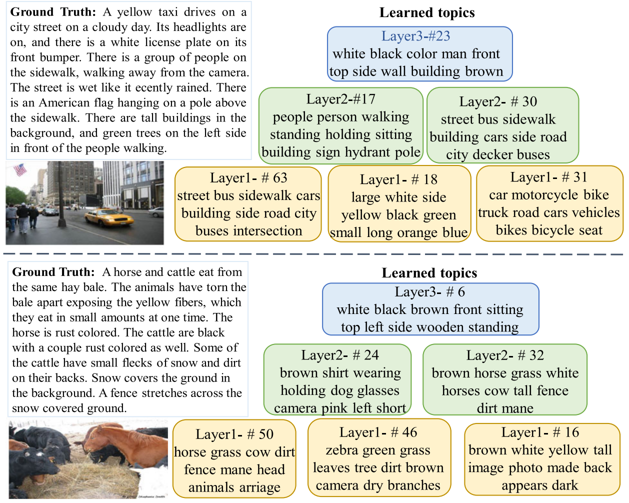
Image	LSTM-ATT	Our VTCM-LSTM	$\mathcal{M}^2$ Transformer	Our VTCM-Transformer	Ground Truth
	there is a train on the tracks . the train is black. the sky is blue with white clouds .	a train is on the tracks . the train is red and black . the train is in motion . the sky is a bright blue with a few clouds .	this is an image of a train engine . the train engine is black and yellow. the letters < unk > is black . there is smoke coming out of the tracks .	a black steam train is driving on the tracks. there is a large cloud of smoke coming out from the train. there are many trees behind the train. The sky is blue with white clouds .	there is a train on the train tracks . there is gray smoke coming out from the front of the train . there is small bushes in front of the train . there are clouds in the sky .
	the bird is standing on the ground . the bird has a black beak . the bird has a long beak and is looking at the camera . the bird has black feathers .	a bird is standing on the beach . the bird is looking down at the water and the water is blue . the bird is a brown color and is standing with a small body on water .	the bird is standing on the beach . the bird has a yellow beak . the bird 's beak is long and yellow . the bird has feathers on it .	a lone bird stands on a beach . the bird has brown feathers and a white body and a yellow beak . the bird has the long legs . the ground under the bird 's head is covered with small waves and branches .	the bird is standing in a small puddle of water . the tip of the birds beak is black . the bird has very bright orange reddish eyes . the birds long legs are a bright yellow color .
	a woman is playing tennis on a tennis court . the woman is wearing a white shirt and white shorts . the woman is holding a racket in her hands . the tennis court is green and green with white lines . there are people sitting in the stands watching the game .	a woman is playing tennis on a tennis court . she is wearing a white shirt and white shorts . there is a woman in a white shirt and black shorts standing on the court . there is a white and white net behind the fence . there are people standing in the stands watching the game .	a woman is in a white shirt and white shorts . she is swinging a tennis racket . there is a tennis ball in front of her .	a woman is running on a tennis court and playing tennis . she is wearing a short sleeve white tank top and skirt . the woman is holding a racket in her hand . the court is surrounded by a fence that is in position green and blue with white painted lines . people are sitting in the stands watching the game .	two people are playing tennis . they are wearing white uniforms . the woman is wearing a tennis skort . she is swinging her racket . the dirt under her is light brown .
	a man is standing on a baseball field . he is wearing a white uniform and a black helmet . the player is holding a white bat in his hand . the field is covered in dirt .	a man is standing on a baseball field . he is wearing a white uniform and a red hat. the field is green . the field is made of dirt and there are a lot of green grass in the background .	a man is standing on a baseball field . he is wearing a white uniform and a red cap on his head . a white ball is soaring towards the man . a green grass can be seen behind the man .	a man in a white uniform is throwing a baseball . he has a red cap on his head . his shoes are black and white . The field is covered with the grass .	a man is throwing a baseball . the baseball is white in color . the man is wearing a red hat . the uniform he is wearing is white .
	a white plate is sitting on a table . the pizza has a white plate and a white tablecloth . the plate is on a white table . a fork and a fork are on the table .	a plate of food is sitting on a table . there is a glass of water and a pizza on the table . there is a white plate with a white napkin on it . there is a glass of wine on the table in front of the plate . there is a fork on the plate .	there is a white plate on the table . there is a silver fork on the plate . there is a silver knife next to the plate .	a round white plate is sitting on a table . the table has a pizza on top of it . the pizza has green garnish on top of it and is cut into half . there are white glasses that are filled with water on the table. a knife and fork are beside the plate of food .	there is a plate sitting on top of a white table . there is a pizza on top of the plate . there is green leaves on top of the pizza . there is a silver fork and a silver knife on top of the pizza next to the plate . there are two glasses on the table behind the plate . there is a slice missing from the pizza .

Fig. 3 Examples for paragraphs generated by LSTM-ATT, the proposed VTCM-LSTM, $\mathcal{M}^2$ Transformer, the proposed VTCM-Transformer, and human-annotated Ground Truth paragraphs on the Stanford dataset. (For better visualization, the novel words are colored in red, the key words of the generated paragraphs and ground truth paragraphs are colored in blue.) (Color figure online)

Under the guidance of visual features and the corresponding interpretable hierarchical topics, our model can thus produce a more relevant description for the image.

Effect of Topics on Paragraph Generation: We hypothesize that the topic information learned from the image visual features can guide the language paragraph generation model to describe the images. Benefiting from the interpretable semantic topics, our model supports the personalized paragraph generation by manipulating the topic information fed into the LSTM-based or Transformer-based LMs. As shown in Fig. 5, for the same image describing a train on the tracks, we make a comparison between the generated captions conditioned on the correct topics predicted by VTCM and the distorted topics. Specifically, the proposed VTCM can infer the image's topic proportion at layer l over global topic $\boldsymbol{\Phi}^l \in \mathbb{R}_+^{K_{l-1} \times K_l}$ as $\boldsymbol{\theta}^l \in \mathbb{R}_+^{K_l}$, which weights the importance of the K_l topics. Therefore, by choosing the index with a maximum value from $\boldsymbol{\theta}^l$ at each layer, we can identify the image's most related topics as #36, #37 and #26 at layers



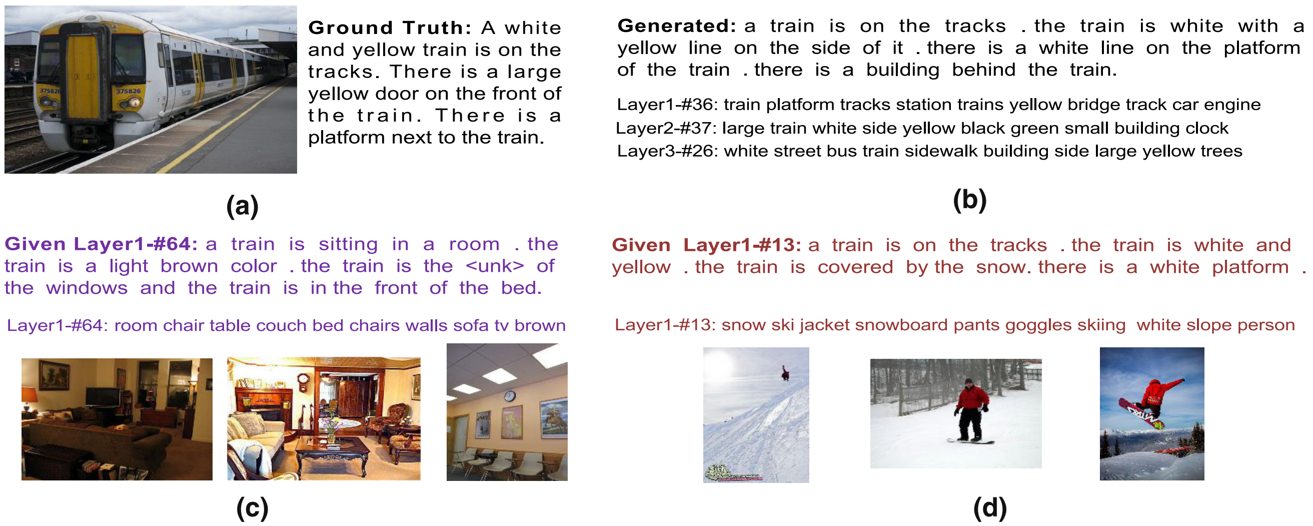
Ground Truth: A yellow taxi drives on a city street on a cloudy day. Its headlights are on, and there is a white license plate on its front bumper. There is a group of people on the sidewalk, walking away from the camera. The street is wet like it ecently rained. There is an American flag hanging on a pole above the sidewalk. There are tall buildings in the background, and green trees on the left side in front of the people walking.

Learned topics
Layer3-#23
white black color man front top side wall building brown
Layer2-#17
people person walking standing holding sitting building sign hydrant pole
Layer2- # 30
street bus sidewalk building cars side road city decker buses
Layer1- # 63
street bus sidewalk cars building side road city buses intersection
Layer1- # 18
large white side yellow black green small long orange blue
Layer1- # 31
car motorcycle bike truck road cars vehicles bikes bicycle seat

Ground Truth: A horse and cattle eat from the same hay bale. The animals have torn the bale apart exposing the yellow fibers, which they eat in small amounts at one time. The horse is rust colored. The cattle are black with a couple rust colored as well. Some of the cattle have small flecks of snow and dirt on their backs. Snow covers the ground in the background. A fence stretches across the snow covered ground.

Learned topics
Layer3- # 6
white black brown front sitting top left side wooden standing
Layer2- # 24
brown shirt wearing holding dog glasses camera pink left short
Layer2- # 32
brown horse grass white horses cow tall fence dirt mane
Layer1- # 50
horse grass cow dirt fence mane head animals arriage
Layer1- # 46
zebra green grass leaves tree dirt brown camera dry branches
Layer1- # 16
brown white yellow tall image photo made back appears dark

Fig. 4 Visualization of the learned topics given the test images, where the top words of each topic at layers 3, 2, and 1 are shown in blue, green, and yellow boxes, respectively. We also present the corresponding ground truth caption for each image, which is not visible at the testing stage (Color figure online)



Ground Truth: A white and yellow train is on the tracks. There is a large yellow door on the front of the train. There is a platform next to the train.

Generated: a train is on the tracks . the train is white with a yellow line on the side of it . there is a white line on the platform of the train . there is a building behind the train.

Layer1-#36: train platform tracks station trains yellow bridge track car engine
Layer2-#37: large train white side yellow black green small building clock
Layer3-#26: white street bus train sidewalk building side large yellow trees

Given Layer1-#64: a train is sitting in a room . the train is a light brown color . the train is the <unk> of the windows and the train is in the front of the bed.

Layer1-#64: room chair table couch bed chairs walls sofa tv brown

Given Layer1-#13: a train is on the tracks . the train is white and yellow . the train is covered by the snow. there is a white platform .

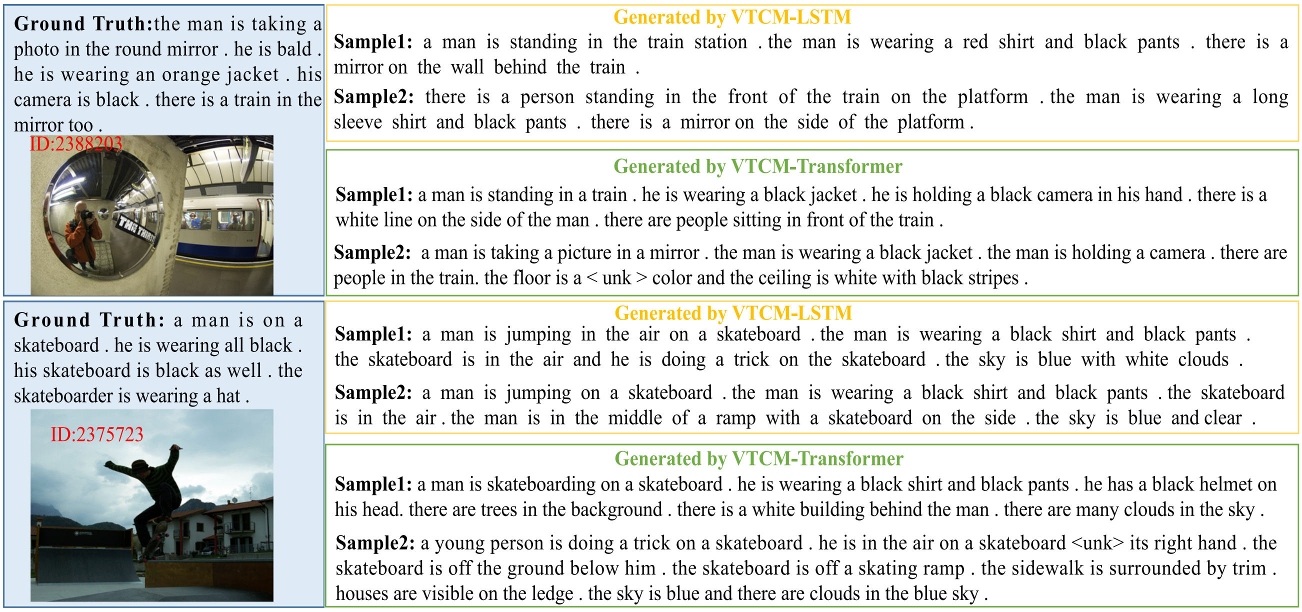
Layer1-#13: snow ski jacket snowboard pants goggles skiing white slope person

Fig. 5 **a** Ground truth caption of an image from Stanford image-paragraph dataset (id = 2349394). **b** The caption generated by the proposed VTCM-LSTM and the corresponding hierarchical topics of the image. The generated caption is able to properly describes the image with the guidance of the correct topic information for paragraph generation. **c–d** The generated captions under the guidance of our selected topics. To see the influence of topic information, we replace the original topics with other randomly-selected topic vectors #64 and #13 at layer 1. It can be seen that the generated captions are influenced by both the input image and the given topics

1, 2 and 3, respectively. Since every topic is a list of representative words, it is easy for users to revise or specify the topic information fed into the paragraph generator by only changing the topic proportion over global topics, where we refer to the designated topic proportion as $\hat{\boldsymbol{\theta}}^{l}$. For example, to specify the k-th topic at layer 1 for the testing image, we can build the one-hot vector $\hat{\boldsymbol{\theta}}^{1}$ as the distorted topic proportion, where $\hat{\theta}^{1}_{k'} = 1$ only if $k' = k$. And we set the $\hat{\boldsymbol{\theta}}^{2:L}$ as zero vectors for simplicity. Clearly, topic #64 at layer 1 is about the room, and the caption about the image is now changed to regarding the train sitting on the room when we set $\hat{\theta}^{1}_{64} = 1$ and feed $\hat{\boldsymbol{\theta}}^{1:L}$ into the paragraph generator. Similarly, topic #13 at layer 1 corresponds to the snow and ski, and the caption now is changed to the train covered by the snow. These observations suggest that the generated captions not only are related to the input image but also reflect what the user wants to emphasize by only changing the topic proportion, making our proposed model controllable.

Diversity: To show the uncertainty in VTCM makes it capable of producing diverse captions while keeping the "main plot" unchanged, we example two descriptions with the same set of inputs, respectively. Different from the bottom row in Fig. 5, here we do not distort the predicted topic proportions from the test image (id=2388203 or id=2375723) but only sample different uniform noises $\boldsymbol{\epsilon}^{1:L}$ to generate $\boldsymbol{\theta}^{1:L}$ via Equation (3), whose posterior parameters transformed from the image features are kept unchanged. Therefore, different noises $\boldsymbol{\epsilon}^{1:L}$ lead to different random topic weight vectors $\boldsymbol{\theta}^{1:L}$, which however share similar weights over global topics and thus similar semantic information. As shown in Fig. 6, our proposed models can generate diverse and coherent paragraphs while ensuring the "big picture" underlying the image does not get lost in the details. The reason behind this might be that our frameworks feed the multi-stochastic-layer latent topic representation $\boldsymbol{\theta}^{1:L}$ of VTCM as the source of randomness to the language generator. Benefiting from the assimilation of multi-stochastic-layer topic information into the language generator, our proposed topic-guided image paragraph captioning systems can guarantee diversity and produce diversified outputs even if there is no specialized module.

The Attention Mechanism in VTCM-LSTM: To evaluate the effectiveness of the attention mechanism in VTCM-LSTM on image captioning, in Fig. 7, we visualize the attended image regions with the biggest attention weight for different words. As we can see, our proposed VTCM-LSTM can reason where the model is focusing on at different time steps. Here we take Image 1 as an example. When predicting "person," the attention module precisely chooses the bounding box covering the main part of the body. While predicting the word "snowboard," our model decides to attend to the surrounding area about the snowboard. It proves that our proposed VTCM-LSTM can capture the alignment between the attended area and the predicted word, which reflects the human intuition during object description.



Ground Truth: the man is taking a photo in the round mirror . he is bald . he is wearing an orange jacket . his camera is black . there is a train in the mirror too .

Generated by VTCM-LSTM

Sample1: a man is standing in the train station . the man is wearing a red shirt and black pants . there is a mirror on the wall behind the train .

Sample2: there is a person standing in the front of the train on the platform . the man is wearing a long sleeve shirt and black pants . there is a mirror on the side of the platform .

Generated by VTCM-Transformer

Sample1: a man is standing in a train . he is wearing a black jacket . he is holding a black camera in his hand . there is a white line on the side of the man . there are people sitting in front of the train .

Sample2: a man is taking a picture in a mirror . the man is wearing a black jacket . the man is holding a camera . there are people in the train. the floor is a < unk > color and the ceiling is white with black stripes .

Ground Truth: a man is on a skateboard . he is wearing all black . his skateboard is black as well . the skateboarder is wearing a hat .

Generated by VTCM-LSTM

Sample1: a man is jumping in the air on a skateboard . the man is wearing a black shirt and black pants . the skateboard is in the air and he is doing a trick on the skateboard . the sky is blue with white clouds .

Sample2: a man is jumping on a skateboard . the man is wearing a black shirt and black pants . the skateboard is in the air . the man is in the middle of a ramp with a skateboard on the side . the sky is blue and clear .

Generated by VTCM-Transformer

Sample1: a man is skateboarding on a skateboard . he is wearing a black shirt and black pants . he has a black helmet on his head. there are trees in the background . there is a white building behind the man . there are many clouds in the sky .

Sample2: a young person is doing a trick on a skateboard . he is in the air on a skateboard <unk> its right hand . the skateboard is off the ground below him . the skateboard is off a skating ramp . the sidewalk is surrounded by trim . houses are visible on the ledge . the sky is blue and there are clouds in the blue sky .

Fig. 6 Different paragraphs generated by the proposed VTCM-LSTM and VTCM-Transformer for two example images from the Stanford image-paragraph dataset (id = 2388203 and 2375723). Given the extracted hierarchical topics $\boldsymbol{\Phi}^{1:L}$ with VTCM from each image, we only sample two different uniform noises $\boldsymbol{\epsilon}^{1:L}$ and produce the corresponding topic weight vectors $\boldsymbol{\theta}^{1:L}$ with randomness using Equation (3), which have similar semantic information and are fed into the language model to depict the input image

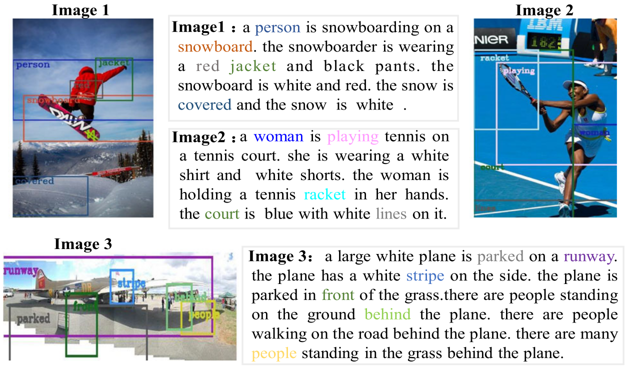


Image1 : a person is snowboarding on a snowboard. the snowboarder is wearing a red jacket and black pants. the snowboard is white and red. the snow is covered and the snow is white .

Image2 : a woman is playing tennis on a tennis court. she is wearing a white shirt and white shorts. the woman is holding a tennis racket in her hands. the court is blue with white lines on it.

Image 3: a large white plane is parked on a runway. the plane has a white stripe on the side. the plane is parked in front of the grass.there are people standing on the ground behind the plane. there are people walking on the road behind the plane. there are many people standing in the grass behind the plane.

Fig. 7 Example of generated captions by VTCM-LSTM showing attended image regions. Different regions and the corresponding words are shown in the same color (Color figure online)

5 Conclusion

We develop a plug-and-play hierarchical-topic-guided image paragraph generation pipeline, which couples a visual extractor with a deep topic model to guide the learning of a language paragraph generation model. As a visual-textual coupling model, the deep topic model can capture the correlations between the image and text at multiple levels of abstraction and learn the semantic topics from images. Serving as the bridge between two modalities, the distilled hierarchical topics are used to guide the caption generation in language model, where we remould both the LSTM-based and Transformer-based language models. Experimental results on the Stanford paragraph dataset show that our proposed models outperform a variety of competing paragraph captioning models, while inferring interpretable hierarchical latent topics and generating semantically coherent paragraphs for the given images.

Acknowledgements B. Chen acknowledges the support of NSFC (U21B2006 and 61771361), Shaanxi Youth Innovation Team Project, the 111 Project (No. B18039) and the Program for Oversea Talent by Chinese Central Government. M. Zhou acknowledges the support of NSF IIS-1812699.

A Appendix

A.1 The Variational Topic Encoder of VTCM

Inspired by Zhang et al. (2018), to approximate the gamma distributed topic weight vector $\boldsymbol{\theta}^l$ with a Weibull distribution, we assume the topic encoder as $q(\boldsymbol{\theta}^l|\overline{\boldsymbol{v}}) = \text{Weibull}(\boldsymbol{k}^l, \boldsymbol{\lambda}^l)$, where the parameters $\boldsymbol{k}^l$ and $\boldsymbol{\lambda}^l$ of $\boldsymbol{\theta}^l$ can be denoted as

$$\boldsymbol{k}^l = \ln[1 + \exp(\mathbf{W}_{hk}^l \boldsymbol{h}^l + \boldsymbol{b}_1^l)], \tag{15}$$

$$\boldsymbol{\lambda}^l = \ln[1 + \exp(\mathbf{W}_{h\lambda}^l \boldsymbol{h}^l + \boldsymbol{b}_2^l)], \tag{16}$$

where $\boldsymbol{h}^l$ are deterministically nonlinearly transformed from the image pooled representation, stated as $\boldsymbol{h}^0 = \overline{\boldsymbol{v}}$ and $\boldsymbol{h}^l = \tanh\left(\mathbf{W}_v^l \boldsymbol{h}^{l-1} + \boldsymbol{b}_v^l\right)$.

A.2 The Gating Unit in VTCM-LSTM

Note the input $\boldsymbol{u}_{j,t}^{l}$ of sentence-level LSTM at layer l combines the topic weight vectors $\boldsymbol{\theta}^{l}$ and hidden output of the sentence-level LSTM $\boldsymbol{h}_{j,t}^{s,l}$ at each time step t. To realize $\boldsymbol{u}_{j,t}^{l} = g\left(\boldsymbol{h}_{j,t}^{s,l}, \boldsymbol{\theta}^{l}\right)$, we adopt a gating unit similar to the gated recurrent unit (GRU) (Cho et al. 2014), defined as

$$\boldsymbol{u}_{j,t}^{l} = \left(1 - z_{j,t}^{l}\right) \odot \boldsymbol{h}_{j,t}^{s,l} + z_{j,t}^{l} \odot \hat{\boldsymbol{h}}_{j,t}^{s,l} \ . \tag{17}$$

where

$$\begin{aligned} z_{j,t}^{l} &= \sigma\left(\mathbf{W}_{z}^{l}\boldsymbol{\theta}^{l} + \mathbf{U}_{z}^{l}\boldsymbol{h}_{j,t}^{s,l} + \boldsymbol{b}_{z}^{l}\right), \\ \boldsymbol{r}_{j,t}^{l} &= \sigma\left(\mathbf{W}_{r}^{l}\boldsymbol{\theta}^{l} + \mathbf{U}_{r}^{l}\boldsymbol{h}_{j,t}^{s,l} + \boldsymbol{b}_{r}^{l}\right), \\ \hat{\boldsymbol{h}}_{j,t}^{s,l} &= \tanh\left(\mathbf{W}_{h}^{l}\boldsymbol{\theta}^{l} + \mathbf{U}_{h}^{l}\left(\boldsymbol{r}_{j,t}^{l} \odot \boldsymbol{h}_{j,t}^{s,l}\right) + \boldsymbol{b}_{h}^{l}\right). \end{aligned} \tag{18}$$

Define $\boldsymbol{u}_{j,t}^{1:L}$ as the concatenation of $\boldsymbol{u}_{j,t}^{l}$ across all layers and $\mathbf{W}_{o}$ as a weight matrix with V rows, the conditional distribution probability $p\left(w_{j,t} \mid w_{j,<t}, Img\right)$ of $w_{j,t}$ becomes

$$p\left(w_{j,t} \mid w_{j,<t}, \boldsymbol{v}_{1:M}, \boldsymbol{\theta}^{1:L}\right) = \operatorname{softmax}\left(\mathbf{W}_{o}\boldsymbol{u}_{j,t}^{1:L}\right). \tag{19}$$

There are two advantages to combine $\boldsymbol{u}_{j,t}^{l}$ at all layers for language generation. First, the combination can enhance representation power because of different statistical properties at different stochastic layers of the deep topic model. Second, owing "skip connections" from all hidden layers to the output, one can reduce the number of processing steps between the bottom of the network and the top, mitigating the "vanishing gradient" problem (Graves et al. 2013).

A.3 Likelihood and Inference of VTCM-Transformer

Given an image Img, we can also represent the paragraph as $\boldsymbol{Y} = \{y_1, ..., y_I\}$, which is suitable for flat language model, such as Transformer-based model. Under the deep topic model (VTCM) and Transformer-based LM, the joint likelihood of the target ground truth paragraph $\boldsymbol{Y}$ of Img and its corresponding BoW count vector $\boldsymbol{d}$ is defined as

$$p\left(\boldsymbol{Y}, \boldsymbol{d} \mid \boldsymbol{\Phi}^{1:L}, \boldsymbol{v}_{1:M}\right) = \int p\left(\boldsymbol{d} \mid \boldsymbol{\Phi}^{1}\boldsymbol{\theta}^{1}\right)\left[\prod_{l=1}^{L} p\left(\boldsymbol{\theta}^{l} \mid \boldsymbol{\Phi}^{l+1}\boldsymbol{\theta}^{l+1}\right)\right] \prod_{i=1}^{I} p\left(y_i \mid y_{<i}, \boldsymbol{v}_{1:M}, \boldsymbol{\theta}^{1:L}\right) d\boldsymbol{\theta}^{1:L}, \tag{20}$$

Since we introduce a variational topic encoder to learn the multi-layer topic weight vectors $\boldsymbol{\theta}^{1:L}$ with the image features $\overline{\boldsymbol{v}}$ as the input. Thus, a lower bound of the log of (20) can be constructed as

$$\begin{aligned} L_{\text{all}} = {} & \mathbb{E}_{q(\boldsymbol{\theta}^{1} \mid \overline{\boldsymbol{v}})}\left[\ln p\left(\boldsymbol{d} \mid \boldsymbol{\Phi}^{1}\boldsymbol{\theta}^{1}\right)\right] \\ & - \sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{\theta}^{l} \mid \overline{\boldsymbol{v}})}\left[\ln \frac{q\left(\boldsymbol{\theta}^{l} \mid \overline{\boldsymbol{v}}\right)}{p\left(\boldsymbol{\theta}^{l} \mid \boldsymbol{\Phi}^{l+1}\boldsymbol{\theta}^{l+1}\right)}\right] \\ & + \sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{\theta}^{l} \mid \overline{\boldsymbol{v}})}\left[\sum_{i=1}^{I} \ln p\left(y_i \mid y_{<i}, \boldsymbol{v}_{1:M}, \boldsymbol{\theta}_{j}^{1:L}\right)\right], \end{aligned} \tag{21}$$

which unites the first two terms primarily responsible for training the topic model component, and the last term for training the Transformer-based LM component. The parameters $\boldsymbol{\Omega}_{\text{TM}}$ of the variational topic encoder and the parameters $\boldsymbol{\Omega}_{\text{Trans}}$ of Transformer-based LM can be jointly updated by maximizing L_{all}. Besides, the global parameters $\boldsymbol{\Phi}^{1:L}$ of the topic decoder can be sampled with TLASGR-MCMC in Cong et al. (2017) and presented below. The training strategy of VTCM-Transformer is similar to that of VTCM-LSTM.

A.4 Inference of Global Parameters $\boldsymbol{\Phi}^{1:L}$ of VTCM

For scale identifiability and ease of inference and interpretation, the Dirichlet prior is placed on each column of $\boldsymbol{\Phi}^{l} \in \mathbb{R}_{+}^{K_{l-1} \times K_{l}}$, which means $0 \leq \Phi_{k',k}^{l} \leq 1$ and $\sum_{k'=1}^{K_{l-1}} \Phi_{k',k}^{l} = 1$. To allow for scalable inference, we apply the topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC algorithm described in Cong et al. (2017); Zhang et al. (2018), which can be used to sample simplex-constrained global parameters in a mini-batch based manner. It improves its sampling efficiency via the use of the Fisher information matrix (FIM), with adaptive step-sizes for the topics at different layers. Here, we discuss how to update the global parameters $\{\boldsymbol{\Phi}^{l}\}_{l=1}^{L}$ of VTCM in detail and give a complete one in Algorithm 1.

Sample the auxiliary counts: This step is about the "upward" pass. For the given mini-batch $\{Img_n, P_n, \boldsymbol{d}_n\}_{n=1}^{N}$ in the training set, $\boldsymbol{d}_n$ is the bag of words (BoW) count vector of paragraph P_n for input image Img_n and $\boldsymbol{\theta}_{n}^{1:L}$ denotes the latent features of the nth image. By transforming standard uniform noises $\boldsymbol{\epsilon}_{n}^{l}$, we can sample $\boldsymbol{\theta}_{n}^{l}$ as

$$\boldsymbol{\theta}_{n}^{l} = \lambda_{n}^{l}\left(-\ln(1 - \boldsymbol{\epsilon}_{n}^{l})\right)^{1/k_{n}^{l}}. \tag{22}$$

Working upward for $l = 1, ..., L$, we can propagate the latent counts x_{vn}^{l} of layer l upward to layer $l + 1$ as

$$A^l_{v1:K_l n} \sim \text{Multi}\left(x^{(l)}_{vn}; \frac{\phi^l_{v1}\theta^l_{1n}}{\sum_{k_l=1}^{K_l}\phi^l_{vk_l}\theta^l_{k_l n}}, \ldots, \frac{\phi^l_{vK_l}\theta^l_{K_l n}}{\sum_{k_l=1}^{K_l}\phi^l_{vk_l}\theta^l_{k_l n}}\right), \tag{23}$$

$$\boldsymbol{m}^{(l)(l+1)}_{kn} = \sum_{v=1}^{K_{l-1}} A^l_{vkn}, \tag{24}$$

$$x^{(l+1)}_{kn} \sim \text{CRT}\left(\boldsymbol{m}^{(l)(l+1)}_{kn}, \boldsymbol{\phi}^{l+1}_{k:}\boldsymbol{\theta}^{l+1}_{n}\right), \tag{25}$$

where $x^1_{vn} = d_{vn}$, $\boldsymbol{d}_n = \{d_{1n}, .., d_{vn}, .., d_{V_c n}\}$, V_c is the size of vocabulary in VTCM, and $x^{(l+1)}_{kn}$ denotes the latent counts at layer $l+1$.

Sample the hierarchical components $\{\boldsymbol{\Phi}^l\}_{l=1}^L$: For $\boldsymbol{\phi}^l_k$, the kth column of the loading matrix $\boldsymbol{\Phi}^l$ of layer l, its sampling can be efficiently realized as

$$\left(\boldsymbol{\phi}^l_k\right)_{q+1} = \Bigg[\left(\boldsymbol{\phi}^l_k\right)_q + \tfrac{\varepsilon_q}{P^l_k}\Big[\left(\rho\tilde{\boldsymbol{A}}^l_{:k\cdot} + \eta^l_0\right) - \left(\rho\tilde{A}^l_{\cdot k\cdot} + K_{l-1}\eta^l_0\right)\left(\boldsymbol{\phi}^l_k\right)_q\Big] + \mathcal{N}\left(0, \frac{2\varepsilon_n}{P^l_k}\left[\text{diag}(\boldsymbol{\phi}^l_k)_q - (\boldsymbol{\phi}^l_k)_q(\boldsymbol{\phi}^l_k)^T_q\right]\right)\Bigg]_\angle, \tag{26}$$

where ε_q denotes the learning rate at the qth iteration, ρ the ratio of the dataset size to the mini-batch size, P^l_k is calculated using the estimated FIM, $\tilde{A}^l_{k'k\cdot} = \sum_{n=1}^N A^l_{k'kn}$, $\tilde{\boldsymbol{A}}^l_{:k\cdot} = \{\tilde{A}^l_{1k\cdot}, \cdots, \tilde{A}^l_{K'k\cdot}\}^T$ and $\tilde{A}^l_{\cdot k\cdot} = \sum_{k'=1}^{K'}\tilde{A}^l_{k'k\cdot}$, $A^l_{k'kn}$ comes from the augmented latent counts A^l in (23), η^l_0 is the prior of $\boldsymbol{\phi}^l_k$, and $[\cdot]_\angle$ denotes a simplex constraint. More details about TLASGR-MCMC for our proposed model can be found in the Equations (18–19) of Cong et al. (2017).

A.5 Additional Experimental Results

To validate the generalizability of our proposed model, we also conducted experiments on the task of generating the radiology reports for the chest X-ray images, which is an important task to apply artificial intelligence to the medical domain. We consider the memory-driven Transformer (M-Transformer) designed for the radiology report generation task (Chen et al. 2020) as our baseline, which introduces a relational memory (RM) to record the information from previous generation processes and a memory-driven conditional layer normalization (MCLN) to incorporate the memory into Transformer. For a fair comparison, we adopt the same implementation for our model; see Chen et al. (2020) for more details. Our experiments are performed on two prevailing radiology report datasets. **IUX-RAY** (Demner-Fushman et al. 2016) is collected by the Indiana University and consists of 7,471 chest X-ray images and 3,955 reports; **MIMIC-CXR** (Johnson et al. 2019) includes 473,057 chest X-ray images and 206,563 reports from 63,478 patients. Following Chen et al. (2020) , we exclude the samples without reports. Note that we can flexibly select the language model for our plug-and-play system, since we pay more attention to assimilating the multi-layer semantic topic weight vectors into the paragraph generator. We here adopt the same Transformer encoder with the M-Transformer and introduce the three-layer semantic topics into its Transformer decoder, where we only add the concatenated topic proportion $\boldsymbol{\theta}^{1:L}$ to the embedding vector of the input token y_{t-1}, which are then embedded to calculate the keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ of the decoder. As summarized in Table 4, our model (VTCM-M-Transformer) outperforms the M-Transformer on METEOR, ROUGE-L, BLEU-1, BLEU-2 and BLEU-3 and is competitive on the BLEU-4. It indicates that the multi-layer semantic topics with VTCM can enhance the radiology report generation despite without designing the complex language model on purpose, proving the generalizability and flexibility of our model. To qualitatively show the effectiveness of our proposed method, we show the reports of two randomly sampled X-ray chest images generated by different methods in Fig. 8, as well as the ground truth reports. Compared with M-Transformer, our VTCM-M-Transformer produces more detailed and coherent reports. For the first image, our VTCM-M-Transformer describes the "heart" and "mediastinal silhouette" in a natural way, while the M-Transformer ignores the "heart". For the second image, the report generated by our model is closer to the ground-truth report, which summarizes the "lungs are hyperexpanded". The above observations show that using the hierarchical semantic topics can enhance the radiology report generation.

Table 4 Comparisons of our proposed VTCM-M-Transformer with M-Transformer on the test sets of IU X-RAY and MIMIC-CXR, where RG-L is ROUGE-L

Data	Method	B-1	B-2	B-3	B-4	M	RG-L
IUX-RAY	M-Transformer	0.470	0.304	0.219	**0.165**	0.187	0.371
	VTCM-M-Transformer	**0.495**	**0.314**	**0.222**	0.162	**0.191**	**0.376**
MIMIC-CXR	M-Transformer	0.353	0.218	0.145	0.103	0.142	0.277
	VTCM-M-Transformer	**0.367**	**0.224**	**0.150**	**0.110**	**0.149**	**0.286**

The best results on the two datasets are marked in bold, respectively

	Ground-truth: both lungs are clear and expanded . heart and mediastinum normal .	M-Transformer: the lungs are clear . the cardiomediastinal silhouette is within normal limits . no pleural effusion is identified .	VTCM-M-Transformer: the lungs are clear of focal airspace disease pneumothorax or pleural effusion . the mediastinal silhouette and heart are within normal limits in size . there are no acute bony findings .
	Ground-truth: the lungs remain hyperexpanded . no xxxx infiltrates or masses . heart and mediastinum are normal .	M-Transformer: the lungs are clear bilaterally . specifically no evidence of focal consolidation pneumothorax or pleural effusion . cardio mediastinal silhouette is unremarkable . visualized osseous structures of the thorax are without acute abnormality .	VTCM-M-Transformer: heart size and mediastinal contour are normal . pulmonary vascularity is normal . lungs are clear . no pleural effusions or pneumothoraces . lungs are hyperexpanded .

Fig. 8 Illustrations of reports from ground-truth, M-Transformer and VTCM-M-Transformer models for two X-ray chest images

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6077–6086).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Burkhardt, S., & Kramer, S. (2019). Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. *Journal of Machine Learning Research, 20*(131), 1–27.

Chatterjee, M., & Schwing, AG .(2018) . Diverse and coherent paragraph generation from images. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 729–744).

Chen, F., Xie, S., Li, X., Li, S., Tang, J., & Wang, T. (2019). What topics do images say: A neural image captioning model with topic representation. In *2019 IEEE international conference on multimedia & expo workshops (ICMEW)* (pp. 447–452), IEEE.

Chen, Z., Song, Y., Chang, T. H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1439–1449).

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734).

Cong, Y., Chen, B., Liu, H., & Zhou, M. (2017). Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *International conference on machine learning* (pp. 864–873), PMLR.

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10578–10587).

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association, 23*(2), 304–310.

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376–380).

Fan, H., Zhu, L., Yang, Y., & Wu, F. (2020). Recurrent attention network with reinforced generator for visual dialog. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16*(3), 1–16.

Fu, K., Jin, J., Cui, R., Sha, F., & Zhang, C. (2017). Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(12), 2321–2334.

Goodfellow, .I, Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649), IEEE.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, 5228–5235.

Guo, D., Chen, B., Lu, R., & Zhou, M.(2020). Recurrent hierarchical topic-guided rnn for language generation. In *International conference on machine learning* (pp. 3810–3821), PMLR.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Huang, L., Wang, W., Chen, J., & Wei, X. Y. (2019). Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4634–4643).

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128–3137).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *2nd International conference on learning representations*.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 317–325).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision, 123*(1), 32–73.

Li, G., Zhu, L., Liu, P., & Yang, Y.(2019). Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8928–8937).

Liang, X., Hu, Z., Zhang, H., Gan, C., & Xing, E. P. (2017). Recurrent topic-transition GAN for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision* (pp. 3362–3371).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

Mao, Y., Zhou, C., Wang, X., & Li, R. (2018). Show and tell more: Topic-oriented multi-sentence image captioning. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 4258–4264).

Melas-Kyriazi, L., Rush, A. M., & Han, G. (2018). Training for diversity in image paragraph captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 757–761).

Miao, Y., Yu, L., & Blunsom, P.(2016). Neural variational inference for text processing. In *International conference on machine learning* (pp. 1727–1736), PMLR.

Ordonez, V., Han, X., Kuznetsova, P., Kulkarni, G., Mitchell, M., Yamaguchi, K., Stratos, K., Goyal, A., Dodge, J., Mensch, A., et al. (2016). Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision, 119*(1), 46–59.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog

Ren, S., He, K., Girshick, R. B., & Sun, J.(2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Shi, Y., Liu, Y., Feng, F., Li, R., Ma, Z., & Wang, X. (2021). S2TD: A tree-structured decoder for image paragraph captioning. In *ACM Multimedia Asia* (pp. 1–7).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd international conference on learning representations*.

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. In *International conference on learning representations*

Tang, J., Wang, J., Li, Z., Fu, J., & Mei, T. (2019). Show, reward, and tell: Adversarial visual story generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15*(2), 1–20.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566–4575).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Wang, J., Pan, Y., Yao, T., Tang, J., & Mei, T. (2019). Convolutional auto-encoding of sentence topics for image paragraph generation. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 940–946).

Wang, J., Tang, J., Yang, M., Bai, X., & Luo, J. (2021). Improving OCR-based image captioning by incorporating geometrical relationship. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1306–1315).

Xu, C., Li, Y., Li, C., Ao, X., Yang, M., & Tian, J. (2020). Interactive key-value memory-augmented attention for image paragraph captioning. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3132–3142).

Xu, C., Yang, M., Ao, X., Shen, Y., Xu, R., & Tian, J. (2021). Retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning. *Knowledge-Based Systems, 214*, 106730.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*| (pp. 2048–2057), PMLR.

Xu, Z., Yang, Y., & Hauptmann, A. G. (2015b). A discriminative CNN video representation for event detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1798–1807).

Yu, N., Hu, X., Song, B., Yang, J., & Zhang, J. (2018). Topic-oriented image captioning based on order-embedding. *IEEE Transactions on Image Processing, 28*(6), 2743–2754.

Zhang, H., Chen, B., Guo, D., & Zhou, M. (2018). WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International conference on learning representations*

Zhang, H., Chen, B., Tian, L., Wang, Z., & Zhou, M. (2020). Variational hetero-encoder randomized generative adversarial networks for joint image-text modeling. In *International conference on learning representations*

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. In *The 30th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 4713–4720).

Zhou, M., Hannah, L., Dunson, D., & Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *Artificial intelligence and statistics* (pp. 1462–1471), PMLR.

Zhou, M., Cong, Y., & Chen, B. (2016). Augmentable gamma belief networks. *Journal of Machine Learning Research, 17*(163), 1–44.

Zhu, L., Fan, H., Luo, Y., Xu, M., & Yang, Y. (2022). Temporal cross-layer correlation mining for action recognition. *IEEE Transactions on Multimedia, 24*, 668–676. https://doi.org/10.1109/TMM.2021.3057503

Zhu, Z., Xue, Z., & Yuan, Z .(2018). Topic-guided attention for image captioning. In *2018 25th IEEE international conference on image processing (ICIP)* (pp. 2615–2619), IEEE.