Uncertainty-aware report generation for chest X-rays by variational topic inference

Najdenkoska, I.; Zhen, X.; Worring, M.; Shao, L.

DOI
10.1016/j.media.2022.102603

Publication date
2022

Document Version
Final published version

Published in
Medical Image Analysis

Medical Image Analysis 82 (2022) 102603

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Uncertainty-aware report generation for chest X-rays by variational topic inference

Ivona Najdenkoska [a,*], Xiantong Zhen [a,b], Marcel Worring [a], Ling Shao [b]

[a] University of Amsterdam, Science Park 904, Amsterdam, The Netherlands
[b] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Dataset link: https://github.com/ivonajdenkoska/variational-xray-report-gen

Keywords:
Chest X-ray
Radiology report
Variational inference
Latent variable model

ABSTRACT

Automating report generation for medical imaging promises to minimize labor and aid diagnosis in clinical practice. Deep learning algorithms have recently been shown to be capable of captioning natural photos. However, doing a similar thing for medical data, is difficult due to the variety in reports written by different radiologists with fluctuating levels of knowledge and experience. Current methods for automatic report generation tend to merely copy one of the training samples in the created report. To tackle this issue, we propose *variational topic inference*, a probabilistic approach for automatic chest X-ray report generation. Specifically, we introduce a probabilistic latent variable model where a latent variable defines a single topic. The topics are inferred in a conditional variational inference framework by aligning vision and language modalities in a latent space, with each topic governing the generation of one sentence in the report. We further adopt a visual attention module that enables the model to attend to different locations in the image while generating the descriptions. We conduct extensive experiments on two benchmarks, namely Indiana U. Chest X-rays and MIMIC-CXR. The results demonstrate that our proposed variational topic inference method can generate reports with novel sentence structure, rather than mere copies of reports used in training, while still achieving comparable performance to state-of-the-art methods in terms of standard language generation criteria.

1. Introduction

Chest X-rays, as one of the most commonly used radiology imaging modalities, are of prime importance for performing diagnosis in clinical routine. For instance, more than 100 million chest X-ray images are obtained annually in the United States alone (Çallı et al., 2021). After this imaging exam is performed, radiologists interpret the chest X-rays and summarize all of the findings into a radiology report. The process of writing such radiology report is known to be time-consuming and tedious even for experienced radiologists.

Recently, automated chest X-ray report generation has come to play an increasingly important role and brings obvious benefits compared to manual report generation. As a result of work overload and staffing shortage, some potential abnormalities may be overlooked or misunderstood, leading to a missed diagnosis. Also, due to the limited experience, radiologists may misinterpret abnormalities in rare diseases. This shows the need for an unbiased diagnosis system. Moreover, there are time-sensitive cases that require immediate reaction and diagnosis. For instance, during the COVID-19 pandemic it is highly desired to quickly and correctly interpret chest X-rays, especially in developing countries. Having accurate automated radiology report generation models can largely reduce workload and speed up clinical practice. More importantly, it could bring clarity to detecting more subtle findings which are not immediately visible to radiologists.

Report generation for chest X-rays can be thought of as transforming visual input into textual output, which is commonly referred to as image captioning. Vinyals et al. (2015), Xu et al. (2015), Lu et al. (2017), Anderson et al. (2018), Cornia et al. (2020), Chen et al. (2021), Zhang et al. (2021). Compared to regular image captioning, medical report generation faces unique challenges, as we need to learn the complex structure of the data and deal with the inherent ambiguity and uncertainty. Naturally, image captioning represents a one-to-many mapping of vision-to-language, as there is not a single best description of an image (Mahajan and Roth, 2020) - there are multiple sentences that can be correct. However, generating multiple correct sentences per image is a difficult problem due to the huge reliance on the exact ground-truth sentences used for training the models. This is also present in the report generation task, resulting in overfitting and producing mere copies of sentences. The diversity modeled in general image

* Corresponding author.
E-mail address: i.najdenkoska@uva.nl (I. Najdenkoska).

https://doi.org/10.1016/j.media.2022.102603
Received 22 February 2022; Received in revised form 11 August 2022; Accepted 24 August 2022
Available online 30 August 2022

captioning, however, is different from the one we aim to model in chest X-ray analysis. The former intents to produce a variety of equally correct image descriptions, whereas the latter should consider a variety of descriptions due to the diversity in data before creating a single best report. Additional challenge compared to regular image captioning, is the fact that X-ray images are much more complex than natural images and require detection of localities, instead of describing the image as a whole. Also, compared to the simple natural image descriptions, the radiology reports demand higher precision of descriptions and more specific medical vocabularies. Last but not least, the lack of enough annotated data, the huge imbalance between normal and abnormal findings, the sensitivity and cost of the data gathering process and the discrepancy in the equipment of different hospitals, make this task even more challenging and versatile. All these characteristics of the problem create unique challenges for automated chest X-ray report generation.

In the latest most successful approaches to chest X-ray report generation (Jing et al., 2018; Li et al., 2018; Liu et al., 2019; Yuan et al., 2019; Xue et al., 2018; Xue and Huang, 2019; Jing et al., 2019; Chen et al., 2020; Lovelace and Mortazavi, 2020; Hou et al., 2021; Liu et al., 2021; You et al., 2021), the neural encoder–decoder architecture is typically used. In particular, a convolutional neural network (CNN) encodes the image into a fixed-size representation and then a language model decodes the representation into a report sentence by sentence. Additional techniques have also been introduced to improve this standard neural architecture, such as incorporating a co-attention mechanism to exploit the relationships between visual features and medical labels (Jing et al., 2018). Other helpful techniques use hierarchical recurrent neural networks (RNNs), such as LSTMs (Hochreiter and Schmidhuber, 1997) to generate multiple sentences and optimize a clinical coherence reward by reinforcement learning to generate reports with high clinical correctness (Liu et al., 2019). To use the information encoded in both frontal and lateral views, (Yuan et al., 2019) explores the fusion of multi-view chest X-rays. Another relevant approach exploits the structure of reports by modeling the relationship between findings and impression sections (Jing et al., 2019). Latest work (Chen et al., 2020; Lovelace and Mortazavi, 2020; Hou et al., 2021; Liu et al., 2021) leverages the Transformer (Vaswani et al., 2017) as a powerful language model to better capture long-term dependencies for sentence generation. Although these deterministic neural encoder–decoder models are the state of the art in terms of benchmark measures, they largely overfit to the training data. They thus produce generic results and are not properly reflecting medical practice. Nevertheless, the basic encoder–decoder architecture forms a good basis for defining chest X-ray report generation models, which we employ in this work.

The process of writing radiology reports is ambiguous and exhibits inherent uncertainty, which is also reflected in the training datasets available for automatic report generation. This uncertainty arises from the fact that the reports are written by radiologists with different levels of expertise, experience and expressive styles. Additionally, it is often difficult to detect small or subtle abnormalities in the images, which results in high inter-observer variability in the chest X-rays analysis (Çallı et al., 2021). Naturally, this yields diversity when several radiologists interpret a single X-ray image into a report. This is almost impossible to be modeled with existing deterministic models, which basically encode the data into fixed-size representations. A more sophisticated approach should learn an approximate distribution of possible high-level patterns in data to improve the generalizability to subtle and unseen cases. Therefore, it is crucial to consider modeling the uncertainty when designing algorithms for report generation to avoid overfitting and achieve generalizability, which is highly necessary in clinical settings.

Probabilistic modeling is well-suited to deal with the uncertainty, diversity, and complex structure of reports (Kohl et al., 2018; Luo and Shakhnarovich, 2020; Mahajan and Roth, 2020). Using stochastic latent variables (Kohl et al., 2018) to represent data prevents to simply compress inputs into fixed-sized deterministic representations. This results in circumventing information loss and allows the holistic characteristics of sentences, such as topic, style, and high-level patterns, to be explicitly modeled (Bowman et al., 2016). Just as important, it is enabling a more diverse and controllable text generation (Wang et al., 2019; Mahajan and Roth, 2020).

To bring these useful characteristics of probabilistic modeling into report generation, we propose variational topic inference (VTI) model, a probabilistic latent variable model for chest X-rays report generation. Particularly, we introduce a set of latent variables, each of which is defined as a topic that governs the sentence generation. The model can be efficiently learned by casting into an optimization objective based on maximizing an evidence lower bound objective (ELBO) (Sohn et al., 2015). During the training process, the topics are inferred from visual and language holistic representations, which are aligned by minimizing the Kullback–Leibler (KL) divergence between them. Moreover, the learning of visual and language holistic representations can be enhanced by using stacks of self-attention layers to learn the important relationships between the features. This essentially means employing a Transformer encoder over the sequence of visual features or language tokens, and pooling special tokens accordingly as holistic representations. At test time the model is able to infer topics from the visual representations only to generate the sentences and maintain coherence between them. In particular, by training the model according to the variational inference framework (Kingma and Welling, 2013) to minimize the KL divergence, we create a latent space where the image and sentence features are aligned. Namely, instead of applying cross-attention as a form of interaction between the multi-modal features, we aim to create an alignment which will capture the high-level patterns of the features. The samples drawn from the latent space, which hold the learned alignment, can be regarded as the topic of a particular sentence related to particular visual features. These latent topics are used to guide the generation module into decoding the sentences. For the generation module, we first adopt two simple LSTMs (Hochreiter and Schmidhuber, 1997), that use the latent topics as memory cell states and are enhanced by visual attention, enabling the model to attend to different local visual regions when generating specific words. As an alternative definition of the generation module we employ a Transformer decoder, which uses the latent topics in the cross-attention module together with the visual features. An overview of the proposed model is given in Fig. 1.

A conference version of this work, which also covers variational topic inference, was published previously (Najdenkoska et al., 2021). The major extension in this work has been made in both methodology and experimental evaluation. We redefine the sentence generator net as a Transformer decoder, which now provides a fully Transformer-based definition of the model. We give a much more thorough derivation of the basic VTI model, augmented with additional architecture figures and algorithms. Moreover, we conduct an additional set of experiments w.r.t the new formulation of the model and we analyze the performance by comparing it to LSTMs. We additionally augment the experiments section by including ablation study and more qualitative evaluation, such as visual attention maps. Last, but not least, we provide a detailed discussion about the imposed diversity in the generated reports and the trade-off between using an LSTM- or Transformer-based decoder for generating medical text.

To summarize, the following is a list of our major contributions: (1) We propose a variational topic inference (VTI) framework to address the chest X-ray report generation problem. (2) We adopt Transformer encoders to aggregate local visual and language features, with each attention head producing a specific topic representation for each sentence, which encourages the generation of a coherent report. (3) We offer two definitions of the sentence generator net, namely an LSTM-based and a Transformer-based model and analyze their performance for chest X-ray report generation. (4) We demonstrate that our method achieves comparable performance to the state of the art on two benchmark datasets under a broad range of evaluation criteria.

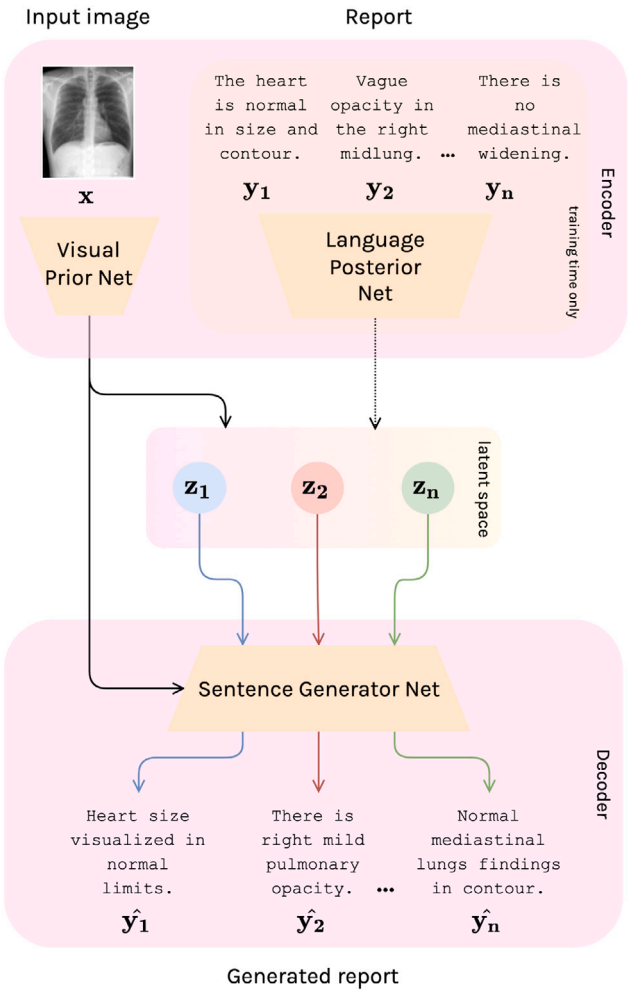

Fig. 1. Overview of the proposed VTI architecture, illustrating the encoder part, which consists of the visual prior net and language posterior net (available only during training time) and the decoder which essentially represents the sentence generator net.

The rest of the paper is organized as follows: In Section 2 we give a detailed overview of the related work in chest X-ray report generation. From there, in Section 3 we formulate the problem in a mathematical manner and introduce the proposed architecture. Section 4 is dedicated to the experimental setup, the used datasets, implementation details and results. A thorough discussion of the experimental findings is given in Section 5 and finally Section 6 concludes the paper.

2. Related work

Report generation aims to produce a few sentences description of a medical image, which is broadly known as image captioning and image paragraph generation in the computer vision field. Arguably, many of the techniques introduced for image captioning are applicable to chest X-ray report generation. We start with regular image captioning and paragraph generation, after which we review related works on chest X-ray report generation.

2.1. Image captioning

Image captioning with a fully neural encoder–decoder model is firstly introduced by Vinyals et al. (2015), where a CNN encoder is used to encode a given image into a fixed-size representation and then an RNN decoder is used to generate a textual description. This base neural encoder–decoder model had many incremental changes over the past years, mainly initiated by the introduction of the attention mechanism (Bahdanau et al., 2015). Starting by incorporating an attention module in the decoder (Xu et al., 2015) and designing adaptive visual attention and attending only to visual words (Lu et al., 2017), many models focus on innovations in the attention module appropriate for image captioning. Around the same time, object detection models emerged, which represented an opportunity for image captioning models to use extracted object features rather than the whole convolutional feature map. Anderson et al. (2018) was the first work that uses object-level information i.e. visual feature maps for each box proposal representing specific salient regions of the image. Instead of operating on a uniform grid of equally-sized image regions, Anderson et al. (2018) calculates the features on object-level regions and then generates a sentence describing the detected objects.

All these encoder–decoder models adopt a common training procedure by minimizing the negative log-likelihood of the current word given the previous ground-truth words, which is essentially the cross-entropy objective. However, the traditional training with cross-entropy suffers from the exposure bias problem imposed by the disparity between the training data distribution and the distribution of its own predicted words. As an alternative, (Rennie et al., 2017) proposes a new training manner for image captioning models, inspired by reinforcement learning. In particular, they optimize the model parameters to maximize an expected reward, which is an evaluation metric, such as CIDEr (Vedantam et al., 2015), used at test time to assess the performance.

As mentioned, the introduction of attention (Bahdanau et al., 2015) brought significant improvements in image captioning models. Therefore, many novel methods focus on improving the definition of attention. For instance, Huang et al. (2019) suggests to extend the conventional attention mechanism to determine the relevance between attention results and queries, denoted as the attended information or the expected useful knowledge. Pan et al. (2020) also propose a modification in the attention block to simultaneously exploit both the spatial and channel-wise attention distributions. Encouraged by the state-of-the-art results of Transformers (Vaswani et al., 2017) in language generation, more recent image captioning models focus on adopting fully Transformer-based models which further boosts their performance. For instance, Herdade et al. (2019) introduce Object Relation Transformer, which incorporates information about the spatial relationship between input detected objects through geometric attention. Cornia et al. (2020) introduce a Transformer-based model augmented with memory slots in the attention module and a meshed connection between the encoder and decoder. Moreover, pre-trained multimodal Transformers, such as Li et al. (2020a,b), Hu et al. (2021) utilize the benefit of self-supervised pre-training with proxy multimodal tasks on large datasets and fine-tuning on downstream tasks, one of which is usually image captioning.

Besides regular image captioning, recent work explores the so-called diverse image captioning. This variant of image captioning tries to replicate the quality and variability of the sentences produced by humans (Stefanini et al., 2021). A natural fit for dealing with diversity in data is probabilistic modeling. For instance, existing work such as Wang et al. (2017) frames image captioning with a conditional variational auto-encoders (CVAEs) with Gaussian Mixture model (GMM) priors to model the diversity in the generated sentences. Mahajan and Roth (2020) also uses conditional variational inference to encode object and contextual information for image-text pairs in the latent space to generate few accurate sentences per image.

Nevertheless, image captioning models focus primarily on generating sentences which describe the image as a whole. On the other hand, chest X-ray report generation typically requires generating multiple coherent sentences per image. There are additional challenges when it comes to generating multiple sentences, which will be addressed in the next section.

2.2. Image paragraph generation

The objective of image paragraph generation extends the one of image captioning, by training models to generate multiple sentences instead of one. This means that the model should be able to identify multiple topics in the image, for example objects of interest and their interaction, and reason about them by generating coherent sentences. Ideally, the sentences in the report should be having a logical transition and without any repetitions.

Among the first works that considered the problem of paragraph generation is Krause et al. (2017), generating entire paragraphs that describe natural images. The main idea is to break the images into semantically meaningful pieces by detecting objects and then to reason about language with a hierarchical RNN decoder. The hierarchical RNN is decomposing the visual features into sentence topics with a sentence-level RNN and generating the sentences with a word-level RNN. This kind of model is extended by Liang et al. (2017) that uses an adversarial framework where the quality of generated sentences produced by paragraph generator, are assessed by adversarial discriminators, which is essentially training a Generative Adversarial Network (GAN)-based model. However, GANs have certain limitations when the goal is to generate sequences of discrete tokens, since they were mainly designed for real-valued, continuous data, such as generating images. Another work Chatterjee and Schwing (2018) proposes a Variational Autoencoder (VAE) formulation of image paragraph generation, which seems like a more appropriate choice compared to GANs. This model also uses hierarchical RNNs and is augmented with so-called coherence vectors and global topic vectors, which help tackling the inherent ambiguity of associating paragraphs with images. The model presented in this work also follows a variational formulation, as an effort to model the implicit ambiguity and uncertainty present the process of chest X-ray reports generation.

2.3. Chest X-ray report generation

Chest X-ray report generation models are largely inspired by image paragraph generation architectures. However, different challenges arise when dealing with chest X-ray image analysis. For instance, the datasets are much smaller than the general ones and are largely imbalanced in terms of normal and abnormal findings, which creates additional challenges for deep learning models. The chest X-ray images are more complicated than natural images and the text exhibits a specific medical vocabulary, much more complex than day-to-day words. Additionally, capturing and describing abnormalities is much more challenging than generating descriptions of normal findings, whereas in general image paragraph generation all sentences are treated equally.

Among the first neural encoder–decoder models that manage to generate diagnostic reports for chest X-rays is Jing et al. (2018). They propose a multi-task learning model which jointly performs prediction of disease labels and generation of reports with a hierarchical LSTM model, similar as Krause et al. (2017). However, it has been noted that this work exhibits some repetitions in the generated reports because their hierarchical LSTM model does not consider the contextual coherence between sentences in reports. To tackle this issue, Xue et al. (2018) proposes a model whose main objective is to achieve intra-paragraph dependency and coherence among the sentences. Another seminal work which follows a similar encoder–decoder architecture and a hierarchical RNN is Yuan et al. (2019). They argue that the fusion of the frontal and lateral views of the chest X-ray is important and they propose various ways to fuse the two views, which slightly improves the performance.

Another line of research argues that chest X-ray report generation can be framed as a retrieval task. It is known that radiologists write reports by following certain patterns and templates, and they adjust certain statements in the templates for each individual case if necessary. Motivated by this, Li et al. (2018) proposes a RL-based model which decides between automatically generating sentences or retrieving specific sentences from the template database, by employing a retrieval policy module. Other work also follows similar retrieval frameworks, such as Sun et al. (2019) proposing a model to retrieve text templates based on the detected abnormalities from the image, and then rewriting the templates according to specific cases, Endo et al. (2021) who proposes a retrieval-based approach using a pre-trained contrastive image-language model and Yang et al. (2021) with their retrieval of reports and sentence-level templates. However, retrieval-based methods experience specific difficulties compared to text generation-based methods. Particularly, they have limitations due to the costly predefined template database in order to retrieve sentences and the explicit construction of templates to determine the patterns embedded in reports.

Last but not least, the introduction of Transformers (Vaswani et al., 2017) is also visible in the recent work on chest X-ray report generation (Chen et al., 2020; Lovelace and Mortazavi, 2020; Hou et al., 2021; Liu et al., 2021; You et al., 2021). For instance, Chen et al. (2020) introduce a memory-driven Transformer with a relational memory designed to record key information of the generation process. Lovelace and Mortazavi (2020) also propose a Transformer-based model trained with the standard language generation and clinical coherence metrics, which specifically addresses the importance of producing clinically correct reports. Moreover, Liu et al. (2021) propose a model which imitates the working process of radiologists, by first examining abnormal regions and assign corresponding disease tags and then relying on the years of prior working experience to write the reports. However, existing work seems to circumvent the issue of uncertainty that is inherently present in the report generation process, which is the main focus of this work.

3. Methodology

3.1. Problem formulation

Given an image $\mathbf{x}$ as input, we aim to generate a report that consists of multiple sentences $\{\mathbf{y}_i\}_{i=1}^{N}$, which are assumed to be conditionally independent. From a probabilistic perspective, we aim to maximize the conditional log-likelihood $\log p_\theta(\mathbf{y}_i|\mathbf{x})$, defined as:

$$\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(\mathbf{y}_i|\mathbf{x}), \tag{1}$$

where θ represents the model parameters and N is the number of sentences in each report. To solve the model, we formulate the report generation as a conditional variational inference problem.

3.2. Variational topic inference

In order to encourage diversity and coherence between the generated sentences in a report, we introduce a set of latent variables $\mathbf{z}$ which are expected to represent topics, each of which governs the generation of one sentence $\mathbf{y}$ in the final report (for brevity we omit the subscript i). By incorporating $\mathbf{z}$ into the conditional probability $p_\theta(\mathbf{y}|\mathbf{x})$, we have the following:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) = \int_z \log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z}) p_\theta(\mathbf{z}|\mathbf{x}) d\mathbf{z}, \tag{2}$$

where $p_\theta(\mathbf{z}|\mathbf{x})$ is the conditional prior distribution. Next, we define a variational posterior $q_\phi(\mathbf{z})$ to approximate the intractable true posterior $p_\theta(\mathbf{z}|\mathbf{y},\mathbf{x})$ by minimizing the KL divergence between them:

$$D_{\mathrm{KL}}[q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})]. \tag{3}$$

Applying Bayes' rule, and using the fact that the KL divergence is non-negative and can be expressed as $D_{\mathrm{KL}}[q \parallel p] = \mathbb{E}[\log(q) - \log(p)]$, we arrive at:

$$D_{\mathrm{KL}}[q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})] = \mathbb{E}\left[\log q_\phi(\mathbf{z}) - \log \frac{p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x})}\right] \geq 0, \tag{4}$$

which gives rise to the evidence lower bound objective (ELBO) of the log-likelihood:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) \geq \mathbb{E}[\log p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})] - D_{\mathrm{KL}}[q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x})] \tag{5}$$

Using the last part of expression (5), it is straightforward to define the ELBO of the model, where the variational posterior $q(\mathbf{z})$ can be designed in various forms to approximate the true posterior.

$$\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi) = \mathbb{E}[\log p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})] - D_{\mathrm{KL}}[q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x})] \tag{6}$$

To leverage the language modality during training, we design the variational posterior as $q_\phi(\mathbf{z}|\mathbf{y})$ conditioned on the ground-truth sentence $\mathbf{y}$. Based on the ELBO, we derive the objective function w.r.t. a report of N sentences as follows:

$$\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi) = \sum_{i=1}^{N}\left[\frac{1}{L}\sum_{\ell=1}^{L}\log p_\theta(\mathbf{y}_i|\mathbf{z}^{(\ell)},\mathbf{x}) - \beta D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{y}_i) \parallel p_\theta(\mathbf{z}|\mathbf{x})]\right], \tag{7}$$

where $\log p_\theta(\mathbf{y}_i|\mathbf{z}^{(\ell)},\mathbf{x})$ is the expected negative reconstruction error in variational auto-encoder parlance and $D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{y}_i) \parallel p_\theta(\mathbf{z}|\mathbf{x})]$ is the KL divergence between the prior and approximate posterior. Additionally, to get a more accurate approximation of the latent topic distribution, we employ Monte Carlo sampling from the conditional distributions which are defined as a simple Gaussian distribution. Particularly, $\mathbf{z}^{(\ell)}$ is the ℓ-th of L Monte Carlo samples, and β is a weighting parameter that controls the behavior of the KL divergence. During training, the samples are drawn from the variational posterior distribution $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z}|\mathbf{y})$, whereas during inference the samples are drawn from the conditional prior distribution $\mathbf{z}^{(l)} \sim p_\theta(\mathbf{z}|\mathbf{x})$.

Note that in the derivation of the ELBO, for the sake of simplicity, we use the notation $\mathbf{x}$ and $\mathbf{y}$ to denote the image and the sentence respectively, but also their holistic feature representations. In the subsequent sections which offer explanation of the learning process, we will define separate notations for the inputs and their holistic representations.

3.3. Learning with neural networks

For efficient optimization, we implement the model with deep neural networks using amortization techniques (Kingma and Welling, 2013). $p_\theta(\mathbf{z}|\mathbf{x})$ and $q_\phi(\mathbf{z}|\mathbf{y})$ are parameterized as fully factorized Gaussian distributions and inferred by multi-layer perceptrons (MLPs), which we refer to as the visual prior net and the language posterior net, respectively. The log-likelihood is implemented as a cross entropy loss based on the output of the sentence generator net and the ground-truth sentence.

Additionally, we employ Transformer encoder and decoder modules to assist the learning of better representations, which are later used in the conditional variational inference networks and the generation of word tokens respectively. In particular, to establish more holistic representations from the encoder, we leverage stacks of multi-head self-attention blocks, which essentially construct a Transformer encoder. We give a formal definition of the used modules, according to Vaswani et al. (2017):

$$\mathtt{MultiHeadAttention}(Q,K,V) = [\mathtt{head}_1;\ldots;\mathtt{head}_h]W_i^O, \tag{8}$$

$$\mathtt{head}_i = \mathtt{Attention}(Q\mathbf{W}_i^Q, K\mathbf{W}_i^K, V\mathbf{W}_i^V), \tag{9}$$

where Q, K and V are the queries, keys and values representations and h is the number of heads in the attention mechanism parlance. $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$, $\mathbf{W}_i^O$ are the linear projections of the queries, keys, values and the output respectively. In addition to the multi-head attention layers, both the encoder and decoder contain a fully connected feed-forward network (FFN), defined as:

$$\mathtt{FFN}(x) = max(0, x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{10}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$ and $\mathbf{b}_2$ are linear projection layers. In the next subsections we will give overview of the building blocks of our model, illustrated on Fig. 1.

3.3.1. Visual prior net

To aggregate the extracted local visual features of input image $\mathbf{x}$ from a pre-trained CNN, into a holistic visual representation, we employ a Transformer encoder. Specifically, the convolutional feature maps are flattened along the spatial dimensions to obtain an array of k local visual features $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k\}$, where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ and d_v is the dimension of a visual feature vector. In particular, to explore the similarity among local features, we adopt the Transformer encoder to encode them into a special visual token $\mathbf{v}_{[\mathtt{IMG}]}$. This special token, initialized to the averaged local visual features, is prepended to the array of k local visual features, as follows: $\mathbf{V} = \{\mathbf{v}_{[\mathtt{IMG}]},\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k\}$. Then the Transformer encoder is defined as stacks of N multi-head attention layers, yielding the following formulation:

$$\mathtt{TrsEncoder}(Q,K,V) = \mathtt{FFN}\big(\mathtt{MultiHeadAttention}(Q,K,V)\big)_N, \tag{11}$$

where Q,K,V are representing the input, which are the local visual features $\mathbf{V}$, meaning that $Q=\mathbf{V}$, $K=\mathbf{V}$ and $V=\mathbf{V}$. Finally, the special token $\mathbf{v}_{[\mathtt{IMG}]}$ is pooled as the holistic representation of the image:

$$\mathbf{v}_{[\mathtt{IMG}]} = \mathtt{TrsEncoder}(\mathbf{V}). \tag{12}$$

This mechanism is essentially inspired by the usage of special tokens in Transformer-based language models (Devlin et al., 2019) and multimodal encoders (Lu et al., 2019). To encourage diversity among topics in a report, we employ a multi-head attention in the Transformer encoder and use each attention head to generate a specific representation for each topic governing the generation of a sentence. The left side of Fig. 2 denotes the visual stream of the encoder i.e. the visual prior net.

3.3.2. Language posterior net

Each sentence $\mathbf{y}$ is represented as a sequence of word tokens including a prepended special language token [SENT]. Each word token is embedded by an embedding matrix W_e, which yields a sequence of n word embeddings $\mathbf{S} = \{\mathbf{e}_{[\mathtt{SENT}]},\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n\}$, where $\mathbf{e}_i \in \mathbb{R}^{d_e}$ and d_e is the dimension of the embedding matrix W_e. Next, a Transformer encoder with positional embeddings encodes the relationships between the word embeddings, which are aggregated into the special token $\mathbf{e}_{[\mathtt{SENT}]}$, pooled as the holistic representation of the sentence:

$$\mathbf{e}_{[\mathtt{SENT}]} = \mathtt{TrsEncoder}(\mathbf{S}), \tag{13}$$

where Q,K,V in Eq. (11) are representing the input, which is the sequence of n word embeddings $\mathbf{S}$, meaning that $Q=\mathbf{S}$, $K=\mathbf{S}$ and $V=\mathbf{S}$.

This net takes ground-truth sentences as input to aid the generation of latent topics during training, thus acting as an additional supervision of the model. The right side of Fig. 2 denotes the language stream of the encoder i.e. the language posterior net.

3.3.3. Conditional variational inference nets

The conditional variational inference nets follow the formulation of standard conditional variational models. The prior distribution $p_\theta(\mathbf{z}|\mathbf{v}_{[\mathtt{IMG}]})$ conditions latent variables on the visual holistic representation i.e. the learned special visual token $\mathbf{v}_{[\mathtt{IMG}]}$. The variational posterior $q_\phi(\mathbf{z}|\mathbf{e}_{[\mathtt{SENT}]})$, which is an approximation to the true posterior, conditions the latent variables on the language holistic representation i.e. the learned special language token $\mathbf{e}_{[\mathtt{SENT}]}$. By minimizing the KL divergence between the conditional prior and variational posterior distributions, as part of the ELBO, the model learns a latent space where the two distributions conditioned on different modalities are eventually aligned. As mentioned before, we assume that the distributions are defined by Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$, where the parameters are estimated by using a simple MLP for the conditional prior and variational posterior accordingly.

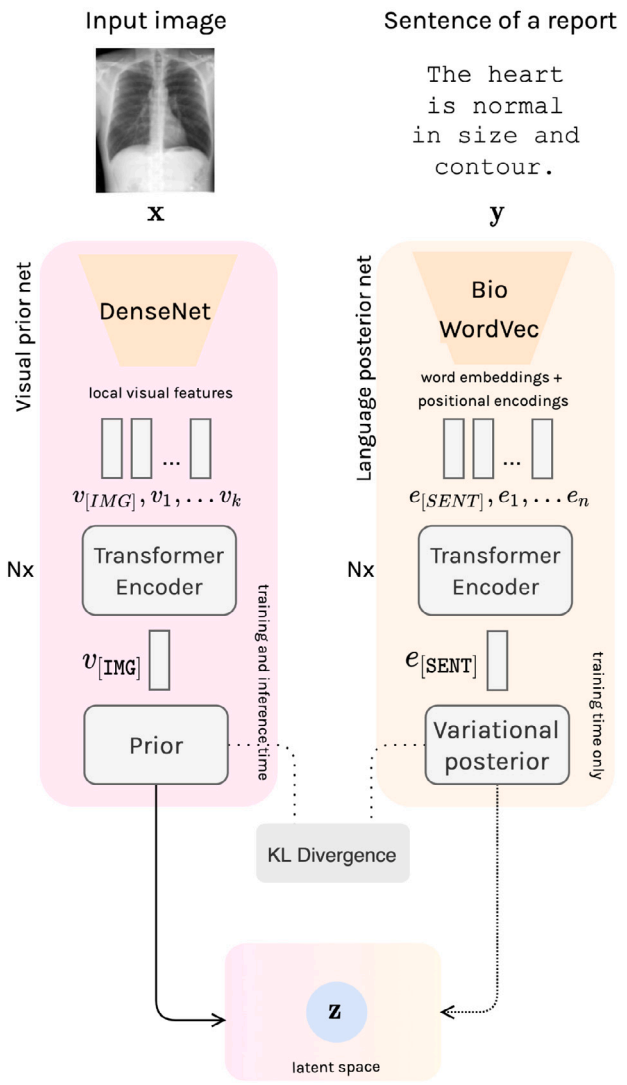

Fig. 2. The encoder of the proposed variational topic inference model, for encoding the chest X-ray image and a single sentence of the report. Separate Transformer encoders with N stacked layers are used to learn a holistic representations of the image and sentence, yielding $v_{[IMG]}$ and $e_{[SENT]}$ respectively. Note that the language stream is only used at training time (indicated by yellow color), when we infer the distributions of latent topics $\mathbf{z}$ from both the visual and language modalities and minimize the KL divergence. At test time, we infer topics from the visual modality only to generate the sentence.

Algorithm 1 Training regime

Input: Training data sample: (Chest X-ray image $\mathbf{x}$, Report $\mathbf{y}$).
Output: Probability distribution of vocabulary words $p_\theta(\hat{\mathbf{y}}|\mathbf{z}^{(\ell)}, \mathbf{x})$ and KL divergence $D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{e}_{[\mathrm{SENT}]})||p_\theta(\mathbf{z}|\mathbf{v}_{[\mathrm{IMG}]})]$
$N \leftarrow$ number of sentences in the report
$L \leftarrow$ number of latent samples
$\mathbf{v}_{[\mathrm{IMG}]}$ =VisualPriorNet($\mathbf{x}$)
$\mathbf{e}_{[\mathrm{SENT}]}$ =LanguagePosteriorNet($\mathbf{y}$)
for $i = 1$ to N **do**
  $\hat{\mathbf{z}} \leftarrow []$
  for $\ell = 1$ to L **do**
    $\mathbf{z}_\ell \sim q_\phi(\mathbf{z}|e^{(i)}_{[\mathrm{SENT}]})$
    Add $\mathbf{z}_\ell$ to $\hat{\mathbf{z}}$
  end for
  $\mathbf{z} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\mathbf{z}}$
  Compute $p_\theta(\hat{\mathbf{y}}_i|\mathbf{z}, \mathbf{x})$ = SentenceGeneratorNet($\mathbf{z}, \mathbf{x}$) and $D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{e}_{[\mathrm{SENT}]})||p_\theta(\mathbf{z}|\mathbf{v}_{[\mathrm{IMG}]})]$ in Eq. (7)
end for

Algorithm 2 Inference regime

Input: Test data sample: (Chest X-ray image $\mathbf{x}$).
Output: Generated report $\hat{\mathbf{y}}$.
$N \leftarrow$ number of sentences in the report
$L \leftarrow$ number of latent samples
$\hat{\mathbf{y}} \leftarrow []$
$\mathbf{v}_{[\mathrm{IMG}]}$ =VisualPriorNet($\mathbf{x}$)
for $i = 1$ to N **do**
  $\hat{\mathbf{z}} \leftarrow []$
  for $\ell = 1$ to L **do**
    $\mathbf{z}_\ell \sim p_\theta(\mathbf{z}|\mathbf{v}^{(i)}_{[\mathrm{IMG}]})$
    Add $\mathbf{z}_\ell$ to $\hat{\mathbf{z}}$
  end for
  $\mathbf{z} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\mathbf{z}}$
  Compute $p_\theta(\hat{\mathbf{y}}_i|\mathbf{z}, \mathbf{x})$ = SentenceGeneratorNet($\mathbf{z}, \mathbf{x}$) and add the most probable $\hat{\mathbf{y}}_i$ to $\hat{\mathbf{y}}$
end for

3.3.4. Sampling latent topics

We perform a pre-defined number of sampling steps L from the conditional variational inference nets and we take the average of the likelihoods, as formulated:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{L}\sum_{l=1}^{L} p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{z}^{(l)}), \tag{14}$$

to obtain a single latent topic. The latent topic essentially represents a high-level pattern, which should steer the sentence generator net into generating the appropriate sentence. To be able to efficiently draw samples and do backpropagation, we use the reparametrization trick as a technique to draw samples from the variational posterior (Kingma and Welling, 2013). In particular, we use the learned μ and σ^2 parameters of the Gaussian distribution to compute $\mathbf{z} = \mu + \sigma^2\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. Algorithm 1 demonstrates the training process when the model uses the sentence holistic representation as a condition of the latent topics, which are drawn from the variational posterior. After being trained, the model is able to infer the latent topics by using the image holistic representation only, as demonstrated in Algorithm 2.

3.3.5. Sentence generator net

The sentences in a report are generated jointly, where the generation of each sentence $\mathbf{y}$ is formulated as $p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{z})$. $\mathbf{y}$ is a sequence of word tokens $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_t$ and it is common to use the joint probability over the tokens to formulate the generation process:

$$p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{z}) = \prod_{t=1}^{T} p_\theta(\mathbf{y}_t|\mathbf{x}, \mathbf{z}, \mathbf{y}_t). \tag{15}$$

The sentence generator net is designed in an autoregressive manner, with two possible implementations explained in the following paragraphs.

LSTM-based sentence generator net. This net contains two consecutive LSTMs according to Anderson et al. (2018), with injected latent topic variables and enhanced by visual attention with hidden states defined by:

$$\mathbf{h}^{(1)}_t = \mathrm{LSTM}^{(1)}(\mathbf{y}_t, \mathbf{h}^{(1)}_{t-1}, \mathbf{c}^{(1)}_{t-1}), \tag{16}$$

where $\mathbf{c}^{(1)}$ is the memory cell state initialized by the latent topic $\mathbf{z}$ for the first time step.

Next, to place focus on different parts of the chest X-ray image while decoding the sentence word by word, we use the concept of visual attention (Xu et al., 2015). The output hidden states $\mathbf{h}^{(1)}_t$ of the first

LSTM at each time step t are used together with the set of k visual features $\mathbf{V}$ to compute the visual attention weights over the image. The sum of both representations is fed into a single-layer neural network followed by a softmax function to generate the attention distribution over the k local visual features of the image:

$$\alpha_t = \mathrm{softmax}\left(\mathbf{w}_a^{\top}\tanh(\mathbf{W}_v\mathbf{V} + \mathbf{W}_h\mathbf{h}_t^{(1)})\right), \tag{17}$$

where $\mathbf{w}_a^T \in \mathbb{R}^k$, $\mathbf{W}_v, \mathbf{W}_h \in \mathbb{R}^{k\times d_h}$ are all learnable parameters. Once the attention distribution α_t is obtained, we can compute the weighted visual representation as follows:

$$\mathbf{v}_a = \sum_{t=0}^{k} \alpha_t \cdot \mathbf{v}_t, \tag{18}$$

which is essentially the aggregated visual representation specific to each word at a given time step t.

The next word in the sequence is predicted by the second LSTM, which takes as input the concatenation of the attended visual representation $\mathbf{v}_a$ and the hidden state $\mathbf{h}_t^{(1)}$ of the first LSTM:

$$\mathbf{h}_t^{(2)} = \mathrm{LSTM}^{(2)}([\mathbf{v}_a; \mathbf{h}_t^{(1)}], \mathbf{h}_{t-1}^{(2)}, \mathbf{c}_{t-1}^{(2)}). \tag{19}$$

Then, the output $\mathbf{h}_t^{(2)}$ of the second LSTM$^{(2)}$ is used to predict the probability distribution p_t of the next word, as in Anderson et al. (2018):

$$p_t = \mathrm{softmax}(\mathbf{W}_p\mathbf{h}_t^{(2)}), \tag{20}$$

where $\mathbf{W}_p \in \mathbb{R}^{d_h\times d_{vocab}}$ is a learnable linear layer that projects $\mathbf{h}_t^{(2)} \in R^{d_h}$ to a probability distribution p_t over the vocabulary of size d_{vocab}. The decoder structure when generating a single sentence of the report is shown on Fig. 3. The model using this decoder is termed VTI-LSTM.

Transformer-based sentence generator net. The definition of the Transformer-based sentence generator net in VTI is following the standard Transformer decoder, whose building blocks are defined in Section 3.3. The input of this module are the encoded local visual features $\mathbf{V}$ of the image as the output of the last Transformer encoder layer. As an additional element of this input set, we include the latent topic $\mathbf{z}$. These features interact with the embeddings of the generated part of the sentence until a given time step t, in N multi-head attention layers, followed by standard fully-connected layers, as shown in:

$$\mathtt{TrsDecoder}(Q, K, V) = \mathtt{FFN}\left(\mathtt{MultiHeadAttention}(Q, K, V)\right)_N, \tag{21}$$

where K and V represent the output of the last encoding layers and Q is the generated output until a given time step t. Then, the output of the last fully-connected layer of the decoder, denoted by $\mathbf{W}_p$ is used to predict the probability distribution p_t of the next word in the sentence:

$$p_t = \mathrm{softmax}\left(\mathbf{W}_p\mathtt{TrsDecoder}([\mathbf{V}; \mathbf{z}], \mathbf{y}_{<t})\right) \tag{22}$$

The decoder structure for generating a single sentence is shown on Fig. 4. This part does not need an explicit visual attention module like the LSTM-based generator, since the Transformer naturally includes a cross-attention module between the encoder and decoder hidden states. The model using this decoder is termed VTI-TRS.

4. Experiments and results

4.1. Datasets and implementation details

We evaluate our VTI model on the Indiana University Chest X-ray collection (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) dataset. Both datasets contain frontal and lateral images, paired with a radiology report. Following standard procedure, images are normalized and resized to 224 × 224, making them appropriate for extracting visual features from a pre-trained DenseNet-121 (Huang et al., 2017). Data entries with missing or incomplete reports are discarded. The impressions and findings sections of the reports are concatenated, lower-cased and tokenized and non-alphabetical words and words that occur less than a pre-defined threshold are filtered out and replaced with a [UNK] token. Shorter sentences and reports are padded to obtain squared batches. After pre-processing, Indiana U. Chest X-ray consists of 3,195 samples, which are split into training, validation and test sets with a ratio of 7:1:2. MIMIC-CXR consists of 218,101 samples and is split according to the official splits.

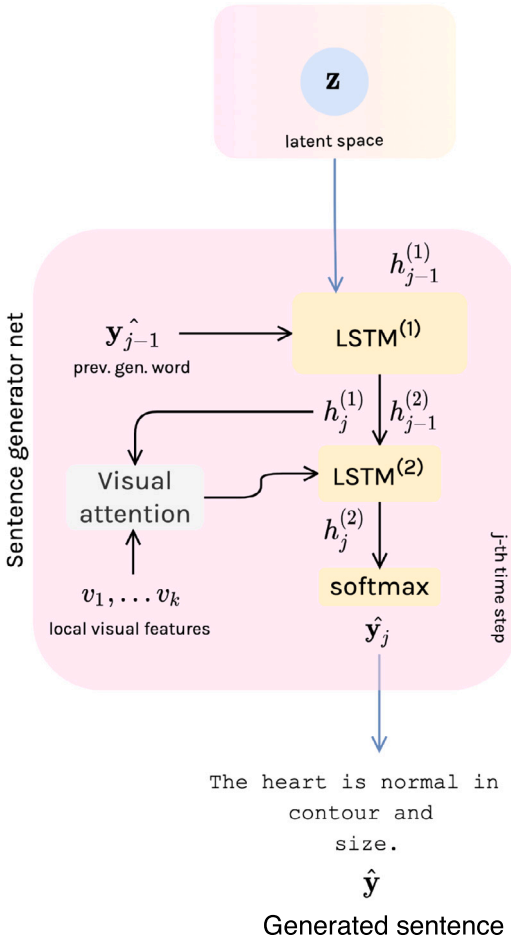

Fig. 3. The LSTM-based decoder of the proposed variational topic inference model. The sample from the latent space z is used as initialization of the memory cell state of the $LSTM^{(1)}$. The hidden state $\mathbf{h}_t^{(1)}$ of the $LSTM^{(1)}$ interacts with the visual features $v_1, \ldots v_k$ in the visual attention module to add the visual information to $LSTM^{(2)}$. The hidden states $\mathbf{h}_t^{(2)}$ of $LSTM^{(2)}$ are projected to the vocabulary space to generate the jth word in the sequence.

The word embeddings are initialized with the pre-trained biomedical embeddings BioWordVec (Zhang et al., 2018), which represent 200-dimensional contextualized vectors. All hyperparameters are set through cross-validation. The linear layers are initialized from a uniform distribution (He et al., 2015) and each one has a hidden dimension of 512, followed by ReLU non-linearity and a dropout with a rate of 0.5. The Transformers in both encoder streams, as well as in the decoder, use a hidden dimension of 512 and have three stacked layers. The model is trained end-to-end on four NVIDIA GTX 1080Ti GPUs using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 3e-05 and early stopping with a patience of five epochs. Last, but not least, we use cyclical annealing technique (Fu et al., 2019) to deal with the notoriously difficult training with KL divergence in the ELBO.

Table 1
Results of VTI model with both LSTM-based and Transformer-based decoders, on Indiana U. Chest X-ray and MIMIC-CXR datasets using the NLG metrics and %Novel. Higher value means better performance for all metrics. The NLG metrics for the other models are cited from the corresponding papers.

Indiana U. X-ray							
Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	%Novel
HRGR-Agent	0.438	0.298	0.208	0.151	–	0.322	–
Clinical-NLG	0.369	0.246	0.171	0.115	–	0.359	–
MM-Att	0.464	0.358	0.270	**0.195**	**0.274**	0.366	–
MvH	0.478	0.334	0.277	0.191	0.265	0.318	–
CMAS-RL	0.464	0.301	0.210	0.154	–	0.362	–
VTI-LSTM (Ours)	**0.493**	**0.360**	**0.291**	0.154	0.218	**0.375**	61.5
Memory-Transformer	0.470	0.304	0.219	0.165	0.187	0.371	–
PPKED	0.483	0.315	0.224	0.168	0.190	0.376	–
AlignTransformer	0.484	0.313	0.225	**0.173**	0.204	0.379	–
Nguyen et al. (MV)	0.476	0.324	0.228	0.164	0.192	0.379	–
VTI-TRS (Ours)	**0.503**	**0.394**	**0.302**	0.170	**0.230**	**0.390**	52.5
MIMIC-CXR							
Clinical-NLG	0.352	0.223	**0.153**	0.104	–	**0.307**	–
VTI-LSTM (Ours)	**0.418**	**0.293**	0.152	**0.109**	**0.177**	0.302	65.2
Memory-Transformer	0.353	0.218	0.145	0.103	0.142	0.277	–
CC-Transformer	0.415	0.272	0.193	0.146	0.159	0.318	–
RATCHET	0.232	–	–	–	0.101	0.240	–
PPKED	0.360	0.224	0.149	0.106	0.149	0.284	–
AlignTransformer	0.378	0.235	0.156	0.112	0.158	0.283	–
Nguyen et al. (MV)	0.451	0.292	**0.201**	**0.144**	0.185	**0.320**	–
VTI-TRS (Ours)	**0.475**	**0.314**	0.196	0.136	**0.191**	0.315	57.3

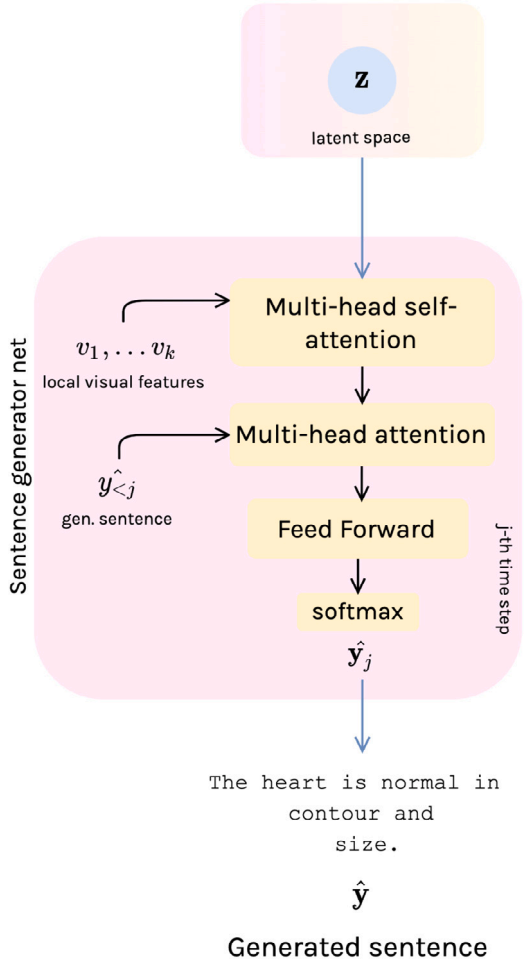

Fig. 4. The TRS-based decoder of the proposed variational topic inference model. The sample from the latent space z is used as input to the multi-head attention layer together with the visual features $v_1, \ldots v_k$. The output of this attention module is then projected to the vocabulary space to generate the jth word in the sequence.

Table 2
Results of VTI model with both LSTM-based and Transformer-based decoders, on MIMIC-CXR dataset using the clinical efficacy metrics. Higher value means better performance for all metrics. The clinical efficacy metrics for the other models are cited from the corresponding papers.

Method	Micro			Macro		
	F1	Precision	Recall	F1	Precision	Recall
Clinical-NLG	–	0.419	**0.360**	–	0.225	**0.209**
VTI-LSTM (Ours)	**0.403**	**0.497**	0.342	**0.210**	**0.350**	0.151
CC-Transformer	0.411	0.475	0.361	0.228	0.333	0.217
Nguyen et al. (MV)	0.533	0.545	0.522	0.347	0.385	**0.347**
VTI-TRS (Ours)	**0.555**	**0.582**	**0.531**	**0.350**	**0.396**	0.312

4.2. Quantitative evaluation

We employ frequently used evaluation metrics for natural language generation (NLG), including BLEU (B) (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009) and ROUGE-L (Lin, 2004). We compare to several other neural network based state-of-the-art methods: HRGR-Agent (Li et al., 2018), Clinical-NLG (Liu et al., 2019), MM-Att (Xue et al., 2018), MvH (Yuan et al., 2019), CMAS-RL (Jing et al., 2019), Memory-Transformer (Chen et al., 2020) and (Nguyen et al., 2021) for Indiana U. X-ray dataset, and Clinical-NLG (Liu et al., 2019), Memory-Transformer (Chen et al., 2020), CC-Transformer (Lovelace and Mortazavi, 2020), PPKED (Liu et al., 2021), AlignTransformer (You et al., 2021), Nguyen et al. (MV) (Nguyen et al., 2021) and RATCHET (Hou et al., 2021) for MIMIC-CXR.

As shown in Table 1, our VTI model achieves comparable performance or yields higher scores in terms of BLEU-1-2-3, ROUGE-L on Indiana U. Chest X-ray dataset and METEOR on MIMIC-CXR, when trained with an LSTM-decoder. Similarly, the model achieves comparable performance when trained with the Transformer-based decoder. In particular, it obtains best results when trained on Indiana U. Chest X-ray dataset and comparable performance or higher scores in terms of BLEU-1-2 and METEOR when trained on MIMIC-CXR. Thanks to the probabilistic nature, our approach prevents the model from generating n-grams with a sentence structure or wording similar to the ground-truth, which is well measured by the NLG metrics. Our approach is able to maintain a better trade-off between accuracy and diversity, which

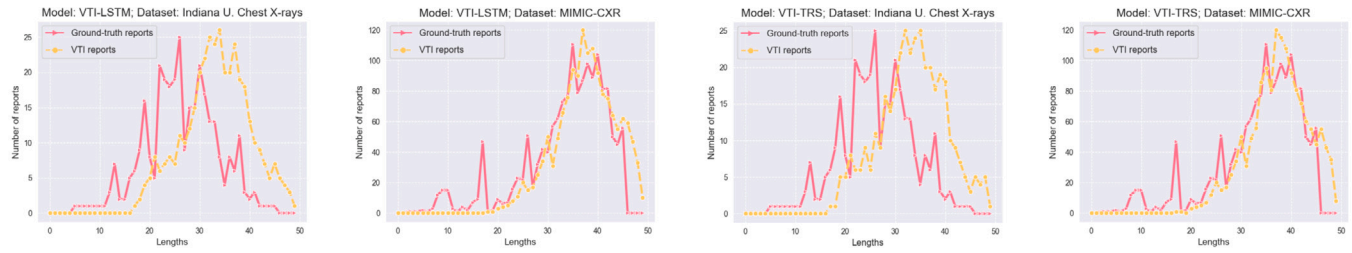

Fig. 5. Length distributions of the ground-truth reports from the Indiana U. Chest X-rays and MIMIC-CXR datasets and the generated reports by VTI model with both the LSTM-based and TRS-based decoders.

is desirable when generating descriptions for images, as pointed out in Luo and Shakhnarovich (2020).

To make a more comprehensive evaluation, we further investigate the clinical coherence and correctness of our proposed model. To do so, we employ clinical efficacy metrics, i.e., precision, recall and F1 score (Liu et al., 2019) to compare the extracted labels by the rule-based CheXpert labeler (Irvin et al., 2019) for the ground-truth and generated reports. For this evaluation, we use the model trained on the MIMIC-CXR dataset, since it contains the 14 disease labels that the CheXpert labeler is trained to extract. The Indiana U. Chest X-ray collection does not contain these particular labels, thus is not used for this evaluation. As shown in Table 2, our model scores higher in precision due to the ability to capture additional information in the chest X-rays, by modeling the diversity of the generated reports. We observe that the model is slightly worse in terms of recall compared to the other models, however, the overall performance, i.e. the F1 score as a harmonic mean between precision and recall is always better than the other baseline models. Another reason for the lower recall is that there might be a mismatch between the ground-truth labels and the ones obtained using the CheXpert labeler from the generated reports. The reason for this is that the generated reports exhibit diversity in terms of sentence structure and topic variability, so the extracted labels will not always match entirely with the ground-truth labels.

Moreover, we plot distributions over the lengths of the generated and ground-truth reports, by VTI with both the LSTM and Transformer decoders, following Chen et al. (2020), in Fig. 5. The generated reports tend to be longer on both datasets, which suggests that more detailed information is captured during decoding. It is worth mentioning that they also follow similar distributions, indicating that our VTI model makes a step closer towards generalizability and is not biased towards a particular dataset or simply replicating the exact ground-truth reports.

To quantify the diversity of the generated results, we incorporate an additional evaluation. Following existing evaluation protocols for measuring diversity in general image captioning literature (Stefanini et al., 2021), we use the %Novel metric as the percentage of generated captions that were not present in the training set. We generate 3 variants of reports per image, as shown in Fig. 7, and we compute their diversity w.r.t the ground-truth report for that image. We repeat this for all samples, and we report the averaged %Novel in Table 1. To better observe the effect of the variational framework on the diversity, we refer to the ablation study in Table 3, where we compute %Novel for the model with and without variational inference module. Omitting this module shows to decrease the diversity, which demonstrates the importance in the generation of novel and diverse reports.

4.3. Ablation analysis

To quantify the impact of the proposed components in the framework, we conduct a few ablation experiments. Firstly, we omit the BioWordVec initialization of the word embedding module, and instead use randomly initialized embedding matrix and retrain the model. It can be observed from Tables 3 and 4 that using a pre-trained embedding matrix, such as BioWordVec, improves the performance and lets the model to be aware of the initial context of the words, which is especially important in medical image analysis, since we deal with very specific and rare vocabulary.

Next, we analyze the effectiveness of the Transformer encoding module in both the visual and language stream. This module gives the holistic representation of the data, which is essentially the special token representation. To ablate this part, we simply take the average of the feature representations as a holistic representation of the image or sentence. From Tables 3 and 4 it can be seen that the addition of Transformer encoder achieves increased performance. This is only an additional empirical proof that self-attention, as a building block of Transformer encoder, helps in learning better aggregation of feature representations independent of their modality.

Furthermore, we quantify the impact of the variational inference part introduced in this paper. In particular, we omit the language stream in the encoding part, since its purpose is to help in the alignment of visual and language features that is essential for sampling of topics. This means that the latent topics are replaced by the special visual token $v_{[\mathrm{IMG}]}$, making the encoder purely deterministic. From Tables 3 and 4, it can be seen that the introduced variational module, together with the language stream improves the overall performance. Additionally, it can be observed that when using the variational topic inference formulation, the diversity of generated reports (%Novel) is higher, which shows the advantage of the newly introduced module.

4.4. Qualitative evaluation

We further examine the performance of the models from a qualitative perspective. Firstly, we observe heat maps of three frontal chest X-rays from the Indiana U. Chest X-ray collection in Fig. 6. These maps are obtained by Grad-CAM (Selvaraju et al., 2017) which is using the gradients flowing into the last convolutional layer to produce a map highlighting important regions in the image. Specifically, they illustrate that the VTI model can focus on relevant image regions while generating the reports. For instance, one of the sentences describing the chest X-ray A is "*Vague opacity in the right midlung*" and as it can be noticed the region around the right midlung is one of the highlighted regions.

Next, for each of the chest X-rays we show three report variants in Fig. 7, in which we draw one topic sample per sentence, demonstrating that different Monte Carlo samples yield variation in the sentence generation process. It can be noticed that the variants describe similar topics with different sentence structures, indicating that the VTI model is aware of more than one correct combination of sentences. Some sentences have variability in their topics owing to the probabilistic modeling. For instance, report 1 for the first chest X-ray image describes the *cardiomediastinal contour* as *normal*, whereas report 2 describes it as *grossly unremarkable*, both with similar semantics. One limitation is that some sentences may have missing words, due to the difficulty of LSTMs to handle long-term dependencies in sequences. This issue is alleviated by using a more powerful language decoder, such as the Transformer decoder.

Nevertheless, VTI can generate reports not necessarily limited to the ground-truth, thus showing awareness to the uncertainty issue and

Table 3
Ablation results on Indiana U. Chest X-ray and MIMIC-CXR using NLG metrics and %Novel. VTI w/o BWV denotes training with a randomly initialized embedding matrix, instead of using the pre-trained BioWordVec. VTI w/o TRS enc means that the holistic representations were obtained by taking the average of the features, instead of using special tokens. VTI w/o var. inf. means that the language stream is omitted and the special visual token $v_{[\mathrm{IMG}]}$ is used as a latent topic. Higher value means better performance for all metrics.

Indiana U. X-ray							
Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	%Novel
VTI-LSTM w/o BWV	0.487	0.352	0.278	0.148	0.211	0.365	–
VTI-LSTM w/o TRS enc	0.435	0.314	0.248	0.121	0.197	0.321	–
VTI-LSTM w/o var. inf.	0.479	0.323	0.289	**0.165**	0.184	0.351	42.5
VTI-LSTM	**0.493**	**0.360**	**0.291**	0.154	**0.218**	**0.375**	**61.5**
VTI-TRS w/o BWV	0.473	0.362	0.281	0.159	0.219	0.373	–
VTI-TRS w/o TRS enc	0.455	0.336	0.261	0.145	0.205	0.344	–
VTI-TRS w/o var. inf.	0.465	0.383	0.285	0.152	0.226	0.378	39.8
VTI-TRS	**0.503**	**0.394**	**0.302**	**0.170**	**0.230**	**0.390**	**52.5**
MIMIC-CXR							
VTI-LSTM w/o BWV	0.405	0.281	0.139	0.102	0.162	0.293	–
VTI-LSTM w/o TRS enc	0.367	0.254	0.119	0.087	0.143	0.278	–
VTI-LSTM w/o var. inf.	0.395	0.285	0.145	0.105	0.164	0.298	50.7
VTI-LSTM	**0.418**	**0.293**	**0.152**	**0.109**	**0.177**	**0.302**	**65.2**
VTI-TRS w/o BWV	0.455	0.305	0.172	0.108	0.178	0.291	–
VTI-TRS w/o TRS enc	0.439	0.293	0.149	0.092	0.152	0.251	–
VTI-TRS w/o var. inf.	0.469	0.305	0.185	0.129	0.188	0.301	41.5
VTI-TRS	**0.475**	**0.314**	**0.196**	**0.136**	**0.191**	**0.315**	**57.3**

Table 4
Ablation results on the MIMIC-CXR dataset using the clinical efficacy metrics (F1, precision and recall). Higher value means better performance for all metrics.

Method	Micro			Macro		
	F1	Precision	Recall	F1	Precision	Recall
VTI-LSTM w/o BWV	0.389	0.479	0.328	0.204	0.343	0.146
VTI-LSTM w/o TRS enc	0.355	0.435	0.301	0.188	0.327	0.132
VTI-LSTM w/o var. inf.	0.370	0.485	0.336	0.198	0.345	0.139
VTI-LSTM	**0.403**	**0.497**	**0.342**	**0.210**	**0.350**	**0.151**
VTI-TRS w/o BWV	0.533	0.519	0.548	0.319	0.361	0.286
VTI-TRS w/o TRS enc	0.497	0.513	0.482	0.300	0.339	0.268
VTI-TRS w/o var. inf.	0.535	0.563	0.513	0.348	0.375	0.305
VTI-TRS	**0.555**	**0.582**	**0.531**	**0.350**	**0.396**	**0.312**

indicating its generalization potential, considered as a major challenge for report generation (Xue and Huang, 2019; Yuan et al., 2019). In clinical scenarios, it is often relevant to have a single best report among a variety. The VTI method is able to produce such a report by combining the most probable sentences given the image, in terms of Bayesian model averaging in a principled way under the probabilistic framework (Kingma and Welling, 2013; Sohn et al., 2015).

To gain more insights into the model, we plot the visual attention maps obtained while generating each word of a given sentence. These maps are produced by the visual attention module of the VTI with an LSTM-based decoder. As it can be observed in Fig. 8, the main finding is that while generating visual words such as *mediastinal*, *silhouette*, *normal*, *cardiomediastinal* and *contours*, the visual attention module puts more weight on certain parts of the image. On the other hand, when generating non-visual words, such as *is*, *the* and *are*, the model does not rely on any particular image region. For cases like this, the model entirely relies on the language model, as also discussed in previous work (Lu et al., 2017).

5. Discussion

5.1. Inherent uncertainty

One of the main objectives of our VTI model is to model the ambiguity and uncertainty inherently present in the interpretation of chest X-rays. In doing so, the VTI model manifests diversity among the generated reports, as can be observed from the qualitative evaluation and Fig. 7. This is a direct effect of the conditional probabilistic design,

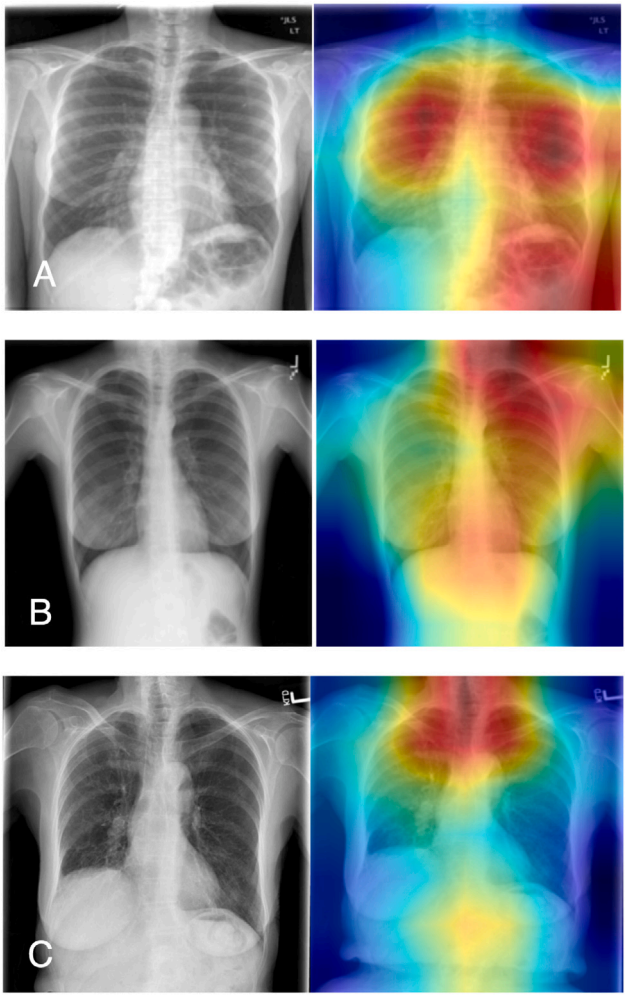

Fig. 6. Examples of frontal chest X-rays and their corresponding heat maps obtained with Grad-CAM (Selvaraju et al., 2017), highlighting relevant image regions for generating a report.

Fig. 7. Examples of chest X-rays, their ground-truth reports and 3 variants of generated reports obtained by our VTI model when trained and evaluated on Indiana U. Chest X-rays. Note that these are the same chest X-rays as the ones in Fig. 6.

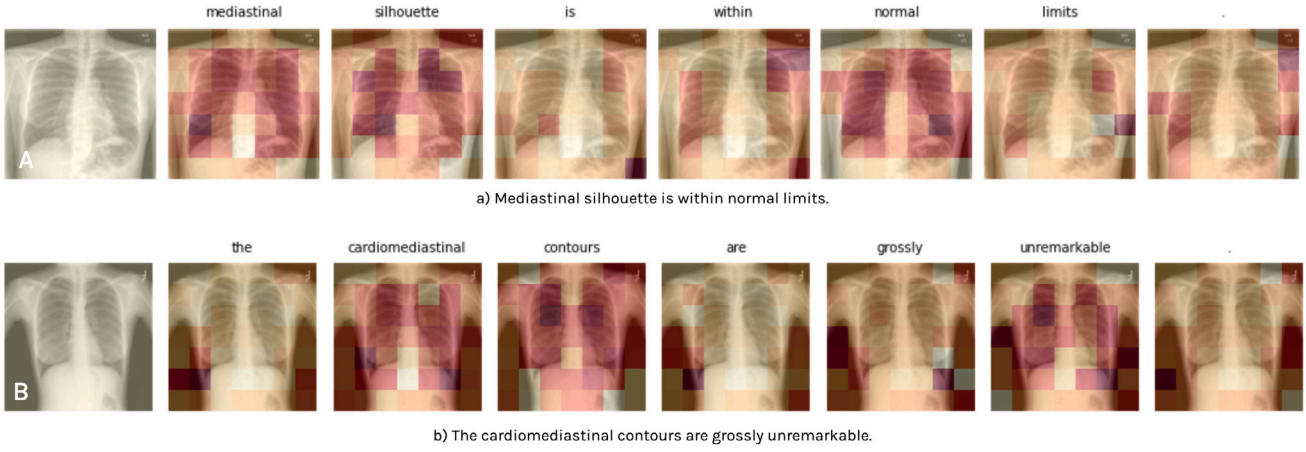

Fig. 8. Examples of the attention maps obtained by the visual attention module of VTI model, while generating each word of a sentence from Indiana U. Chest X-rays.

which we deem important as the data we are dealing with is inherently diverse.

The probabilistic modeling is motivated by the observation that sentences in a report can be diverse in terms of sentence structures and styles because they are written by radiologists with a different experience or expertise. Therefore, the report could exhibit large uncertainty in terms of the wording and semantic structure of the whole sentences. By framing report generation with a Bayesian inference formalism, our model manages to capture this uncertainty to better handle the diversity in the training data. Our probabilistic model has the innate ability to avoid overfitting to training data, therefore potentially offering a better generalization performance of report generation. Therefore, the model is aware that there are multiple ways of describing a given diagnosis, which is visible in the variations in the sentence structures. Moreover, in clinical settings it is a common scenario to have different opinions about a given chest X-ray image by different radiologists. In a similar way our method offers the possibility to produce alternative reports, instead of only one.

Finally, a well-known challenge in chest X-ray report generation is the evaluation of the clinical correctness of reports. Current commonly used evaluation metrics for natural language generation, such as BLEU, METEOR and ROUGE-L, are evaluating the structure of the sentences and are not taking into account whether a diagnosis was correctly detected. We argue that having a variety of reports to present to a

radiologist to choose from can create a higher chance of capturing a correct diagnosis.

5.2. LSTM- vs transformer-based decoder

This work proposes a model with flexible design of its components, meaning that the decoder can be implemented as any language model. We implement the decoder with two instantiations of such methods, namely an LSTM-based (Anderson et al., 2018) and a Transformer-based (Vaswani et al., 2017) decoder.

Initially, we hypothesized that the Transformer will produce superior results compared to the LSTM, since it is entitled as state-of-the-art when it comes to natural language processing tasks, such as generation of longer text sequences. When observing the results in Tables 1 and 2, it can be noticed that the Transformer-based decoder indeed outperforms the LSTM-based decoder, but by a small margin. A justification for this small margin can be the fact that chest X-ray report generation is a very specific task and it requires close-ended descriptions which must be clinically plausible. Transformers, however, shine the most when it comes to generating general and open-ended text, such as stories, poems, news or even coding (Brown et al., 2020). Another point is that Transformers benefit from large datasets, and available chest X-ray report datasets are not large enough. Therefore, training a Transformer still requires lots of engineering and tuning of hyperparameters. Also, it needs longer time to converge and it is computationally expensive, which was also observed as an issue in our experiments. However, generating medical text with a Transformer is still under-explored, compared to LSTMs. The latter deals with the notoriously difficult problem of learning long term dependencies, resulting in missing words or repetitions, which is not the case with the Transformer.

To conclude, the proposed variational topic inference is a general probabilistic modeling framework which can be implemented with different decoder architectures for sentence generation. We have implemented both LSTM and Transformer based decoders, since both have their own advantages and disadvantages. We found that LSTMs can still be competitive with Transformers when it comes to generating specific medical text, especially without too much memory and computational cost. In future, the Transformer will probably be the model of choice for this task, since it is superior in generating longer sequences of text, which is highly desirable in chest X-ray report generation.

6. Conclusion

In this paper, we present a probabilistic approach for dealing with the uncertainty and ambiguity that occur in the chest X-ray interpretation process. We frame the task as a variational inference problem which gives rise to a novel, theoretically sound probabilistic method for dealing with common report generation challenges, including generic and repetitive words. We define topics as latent variables and align the vision and language modalities in the latent space. These latent variables are used by a sentence generator net, which can be implemented by either an LSTM or a Transformer, to guide the sentence generation process. We conduct extensive experiments on two benchmark datasets, Indiana University Chest X-rays and MIMIC-CXR. This approach demonstrates to be able to deal with the inherent uncertainty and ambiguity in the data, achieving competitive performance with deterministic architectures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code available on: https://github.com/ivonajdenkoska/variational-xray-report-gen.

Acknowledgments

This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

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: IEEE Conference on Computer Vision and Pattern Recognition.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. CoRR, arXiv:1409.0473.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S., 2016. Generating sentences from a continuous space. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. pp. 10–21.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K., 2021. Deep learning for chest X-ray analysis: A survey. Med. Image Anal. 72, 102125.

Chatterjee, M., Schwing, A.G., 2018. Diverse and coherent paragraph generation from images. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 729–744.

Chen, L., Jiang, Z., Xiao, J., Liu, W., 2021. Human-like controllable image captioning with verb-specific semantic roles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16846–16856.

Chen, Z., Song, Y., Chang, T.-H., Wan, X., 2020. Generating radiology reports via memory-driven transformer. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

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., Rosenman, M., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G., McDonald, C., 2016. Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. 304–310.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics.

Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P., 2021. Retrieval-based chest X-Ray report generation using a pre-trained contrastive language-image model. In: Proceedings of Machine Learning for Health. In: Proceedings of Machine Learning Research, vol. 158, PMLR, pp. 209–219.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., Carin, L., 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In: North American Chapter of the Association for Computational Linguistics. pp. 240–250.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: IEEE International Conference on Computer Vision. pp. 1026–1034.

Herdade, S., Kappeler, A., Boakye, K., Soares, J., 2019. Image captioning: Transforming objects into words. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc..

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hou, B., Kaissis, G., Summers, R.M., Kainz, B., 2021. RATCHET: Medical transformer for chest X-ray diagnosis and reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 293–303.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L., 2021. Scaling up vision-language pre-training for image captioning. arXiv preprint arXiv:2111.12233.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, L., Wang, W., Chen, J., Wei, X.-Y., 2019. Attention on attention for image captioning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: 33rd AAAI Conference on Artificial Intelligence.

Jing, B., Wang, Z., Xing, E., 2019. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 6570–6580.

Jing, B., Xie, P., Xing, E., 2018. On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

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.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kohl, S.A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Eslami, S., Rezende, D.J., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images. arXiv preprint arXiv:1806.05034.

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.

Lavie, A., Denkowski, M.J., 2009. The meteor metric for automatic evaluation of machine translation. Mach. Transl. 105–115.

Li, Y., Liang, X., Hu, Z., Xing, E.P., 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems, vol. 31.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al., 2020a. Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. Springer, pp. 121–137.

Li, L.H., You, H., Wang, Z., Zareian, A., Chang, S.-F., Chang, K.-W., 2020b. Unsupervised vision-and-language pre-training without parallel images and captions. arXiv preprint arXiv:2010.12831.

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.

Lin, C.-Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries. Association for Computational Linguistics (ACL).

Liu, G., Hsu, T.-M.H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., Ghassemi, M., 2019. Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269.

Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y., 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13753–13762.

Lovelace, J., Mortazavi, B., 2020. Learning to generate clinically coherent chest X-Ray reports. In: Findings of the Association for Computational Linguistics: EMNLP. pp. 1235–1243.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural Inf. Process. Syst. 32.

Lu, J., Xiong, C., Parikh, D., Socher, R., 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 375–383.

Luo, R., Shakhnarovich, G., 2020. Analysis of diversity-accuracy tradeoff in image captioning.

Mahajan, S., Roth, S., 2020. Diverse image captioning with context-object split latent spaces. In: Advances in Neural Information Processing Systems (NeurIPS).

Najdenkoska, I., Zhen, X., Worring, M., Shao, L., 2021. Variational topic inference for chest X-Ray report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 625–635.

Nguyen, H., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., Cheng, L., 2021. Automated generation of accurate & fluent medical X-ray reports. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 3552–3569.

Pan, Y., Yao, T., Li, Y., Mei, T., 2020. X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10971–10980.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: A method for automatic evaluation of machine translation. In: Association for Computational Linguistics. pp. 311–318.

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision. pp. 618–626.

Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R., 2021. From show to tell: A survey on image captioning. arXiv preprint arXiv:2107.06912.

Sun, L., Wang, W., Li, J., Lin, J., 2019. Study on medical image report generation based on improved encoding-decoding method. In: International Conference on Intelligent Computing. Springer, pp. 686–696.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.

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: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164.

Wang, W., Gan, Z., Xu, H., Zhang, R., Wang, G., Shen, D., Chen, C., Carin, L., 2019. Topic-guided variational auto-encoder for text generation. North Am. Chap. Assoc. Comput. Linguist..

Wang, L., Schwing, A., Lazebnik, S., 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. Adv. Neural Inf. Process. Syst. 30.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057.

Xue, Y., Huang, X., 2019. Improved disease classification in chest X-Rays with transferred features from report generation. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (Eds.), Information Processing in Medical Imaging. pp. 125–138.

Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X., 2018. Multimodal recurrent model with attention for automated radiology report generation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 457–466.

Yang, X., Ye, M., You, Q., Ma, F., 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics.

You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X., 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer International Publishing, Cham, pp. 72–82.

Yuan, J., Liao, H., Luo, R., Luo, J., 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z., 2018. BioWordVec: Improving biomedical word embeddings with subword information and MeSH ontology.

Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R., 2021. RSTNet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15465–15474.