scientific reports

OPEN

Medical image captioning via generative pretrained transformers

Alexander Selivanov[1,2,4], Oleg Y. Rogov[1,4], Daniil Chesakov[1,3], Artem Shelmanov[1,3], Irina Fedulova[2] & Dmitry V. Dylov[1]✉

The proposed model for automatic clinical image caption generation combines the analysis of radiological scans with structured patient information from the textual records. It uses two language models, the Show-Attend-Tell and the GPT-3, to generate comprehensive and descriptive radiology records. The generated textual summary contains essential information about pathologies found, their location, along with the 2D heatmaps that localize each pathology on the scans. The model has been tested on two medical datasets, the Open-I, MIMIC-CXR, and the general-purpose MS-COCO, and the results measured with natural language assessment metrics demonstrated its efficient applicability to chest X-ray image captioning.

Medical imaging is indispensable in the current diagnostic workflows. Out of the plethora of existing imaging modalities, X-ray remains one of the most widely-used visualization methods in many hospitals around the world, because it is inexpensive and easily accessible[1]. Analyzing and interpreting X-ray images is especially crucial for diagnosing and monitoring a wide range of lung diseases, including pneumonia[2], pneumothorax[3], and COVID-19 complications[4].

Today, the generation of a free-text description based on clinical radiography results has become a convenient tool in clinical practice[5]. Having to study approximately 100 X-rays daily[5], radiologists are overloaded by the necessity to report their observations in writing, a tedious and time-consuming task that requires a deep domain-specific knowledge. The typical manual annotation overload can lead to several problems, such as missed findings, inconsistent quantification, and delay of a patient's stay in the hospital, which brings increased costs for the treatment. Among all, the qualification of radiologists as far as the correct diagnosis establishing should be stated as major problems.

In the COVID-19 era, there is a higher need for robust image captioning[5–7] framework. Thus, many healthcare systems outsource the medical image analysis task. Automatic generation of chest X-ray medical reports using deep learning can assist and accelerate the diagnosis establishing process followed by clinicians. Providing automated support for this task has the potential to ease clinical workflows and improve both care quality and standardization. For that, we propose to adapt powerful models from non-medical domain.

Medical background. Radiology is the medical discipline that uses medical imaging to diagnose and treat diseases. Today, radiology actively implements new artificial intelligence approaches[8–10]. There are three types of radiologists—diagnostic radiologists, interventional radiologists and radiation oncologists. They all use medical imaging procedures such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), nuclear medicine, positron emission tomography (PET) and ultrasound. Diagnostic radiologists interpret and report on images resulted from imaging procedures, diagnose the cause of patient's symptoms, recommend treatment and offer additional clinical tests. They specialize on different parts of human body—breast imaging (mammograms), cardiovascular radiology (heart and circulatory system), chest radiology (heart and lungs), gastrointestinal radiology (stomach, intestines and abdomen), etc. Interventional radiologists use radiology images to perform clinical procedures with minimally invasive techniques. They are often involved in treating cancer, heart diseases, stroke, blockages in the arteries and veins, fibroids in the uterus, back pains, liver and kidney problems.

[1]Skolkovo Institute of Science and Technology, Bolshoy blvd., 30/1, Moscow 121205, Russia. [2]Philips (Russia), Skolkovo Technopark 42, Building 1, Bolshoi Boulevard, Moscow 121205, Russia. [3]AIRI, Kutuzovsky Ave, 32 bld. 1, Moscow 121170, Russia. [4]These authors contributed equally: Alexander Selivanov and Oleg Y. Rogov. ✉email: d.dylov@skoltech.ru

Technical background. Because image captioning is a multimodal problem, it draws a significant attention of both computer vision and natural language processing communities. The latest surveys in the medical image captioning task[5,11] offer a detailed description of domain knowledge from radiology and deep learning. The first architectures to address this problem were CNN-RNN models from[12,13]. However, the latter shows satisfactory results only on the single-pathology tasks.

With the new concept of attention approach[14], more papers have begun to use visual attention[15–17], being the first to use attention on medical images. The authors of[15] presented a model that can fix its attention on salient objects while generating the corresponding words in the output sequence. Shortly after the visual-attention concept was exposed, text-attention was introduced by authors of TieNet—a framework that generates natural reports[18–20] for the Chest-Xray dataset[21]. They used both semantic and visual attention, that allowed them to get high natural language generation (NLG) metrics on medical datasets. It was trained for solving several tasks such as classification, localization, and text generation. It used a non-hierarchical CNN-LSTM[22] approach together with the attention to semantic and visual features, as it allowed to overperform the current state-of-the-art results. In the[23], bone fracture X-ray reports were generated by identifying image features and filling text templates. The authors of[20] suggested a multi-task framework, that can both predict tags and generate texts using co-attention. This model is still not sufficient for producing accurate diagnosis from X-rays as the produced texts still contained repeated sentences due to a lack of contextual coherence in the hierarchical models. The authors of[24] took advantage of a sentence-level attention mechanism in a late fusion fashion. They took advantage of the multi-view images from both frontal and lateral view angles from the Open-I dataset[25].

The authors of[26] proposed to utilize a pre-constructed knowledge graph embedding module (extracted from the Open-I images using Chexnet models[27]) on multiple disease findings to assist the report generation process. The authors of[28] exposed an anomaly detection method for detecting abnormalities on chest X-rays with deep perceptual autoencoders. The authors of[29] first generated topics for sentences using reinforcement learning (RL) followed by the word decoder sequence generation from the topic with attention to the original images. RL was used for tuning to optimize readability. We solve this problem in a simpler method without losing in quality. To extract topics, we use the NegBio labeller[21,30], which provides topics from clinical reports. We add these topics to the beginning of the medical report, for our model to understand where exactly the text should be generated.

The work[31] focuses on reporting abnormal findings on radiology images. The proposed method learns conditional visual-semantic embeddings in radiology images; and the reports are further used to measure the similarity between the image regions and the medical reports. This by optimizing a triplet ranking loss. The authors of[32] developed an algorithm that learns a description of findings from images and uses their pattern of occurrences to retrieve and customize similar reports from a large report database. The work in[33] proposed a Contrast Induced Attention Network (CIA-Net), using contrastive learning on the aligned positive and negative samples for the disease localization on the chest X-ray images. The work in[34] studies the cross-domain performance, agreement between models, and model representations for X-rays diagnostic prediction tasks. The authors test for concept similarity by regularizing a network to group tasks across multiple datasets together and observe variation across the tasks. The model in[22] generates a short textual summary with essential information on the found pathologies along with their location and severity. The model is trained on only 2% of the MIMIC-CXR dataset, and generates short reports. Although, in this work, we train on the whole MIMIC-CXR and generate a full-text report.

The authors of[35–39] attempted to use transformer-based models as decoders in the image captioning domain[22]. The work[38] affirmed to have generated radiology reports through the custom transformer with additional memory-driven unit. Another model was introduced in[39] where encoder detects regions of interest via a bottom-up attention module and extracts top-down visual features. In this study, the decoder is presented as a custom transformer. For example, the paper in[36] proposes an approach called "pseudo self-attention". Its main idea is to incorporate the conditioning input as a pseudo history to a pretrained transformer. They add a new key and value weights in the self-attention module to be projected onto the decoder's self-attention space, while[37] focuses on visual and weighted semantic features.

Contributions. The contributions of this paper are the following:

- We introduce a new architecture for image captioning, based on a combination of two language models with image-attention (SAT) and text-attention (GPT-3), outperforming current state-of-the-art models
- We introduce a new preprocessing pipeline for radiology reports that allows to get higher NLG metrics
- We perform extensive experiments to show the capability of the proposed method
- Finally, we contribute to deep learning community by training two language models on a large dataset MIMIC-CXR

The rest of the paper is organized as follows: section "Methods" describes the architecture of two language models separately, section "Proposed architecture" provides the description of the proposed approach, section "Experiments" describes the data and the computing, the last sections compare the results and conclude the paper.

Methods

Show attend and tell. Show Attend and Tell (SAT)[15] is an attention-based image caption generation neural net. An attention-based technique allows to get well interpretable results, which can be utilized by radiologist to ensure their findings on X-ray. By including attention, the module gives the advantage to visualize where exactly the model 'sees' the specific pathology. SAT consists of three blocks: Encoder, Attention module, and Decoder.

It takes an image, encodes it, attends each part of the image, and generates an L-length caption $\mathbf{z}$, an encoded sequence of words from the W-length vocabulary:

$$\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_L\}, \quad \mathbf{z}_i \in \mathbb{R}^{W_{SAT}} \tag{1}$$

Encoder. Encoder is a convolutional neural network (CNN). It encodes an image and outputs a set of C vectors, each of which is a D-dimensional representation of the image corresponding part:

$$\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_C\}, \quad \mathbf{a}_i \in \mathbb{R}^{D \times D} \tag{2}$$

Here, C represents the number of channels in the output of the encoder. It depends on the used type of the encoder: 1024 for DenseNet-121[40], 512 for VGG-16[41], 2048 for InceptionV3[42] and ResNet-101[43]. D is a configurable parameter representing the encoded vectors size. Features are extracted from the lower convolutional layer prior to the fully connected layers, and are being passed through the Adaptive Average Pooling layer. This allows the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors.

Decoder with attention module. The decoder is implemented as an LSTM neural network[44]. It produces a caption by generating one word at every time step conditioned by the attention (context) vector, the previous hidden state and the previously generated words. The LSTM can be represented as the following set of equations:

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{E}\mathbf{z}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{a}}_t \end{pmatrix} \tag{3}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{4}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \tag{5}$$

Vectors $\mathbf{i}_t, \mathbf{f}_t, \mathbf{c}_t, \mathbf{o}_t, \mathbf{h}_t$ represent the input/update gate activation vector, forgetting gate activation vector, memory or cell state vector, while outputting gate activation vector and hidden state of the LSTM respectively. $T_{s,t}$ is an affine transformation, such that $\mathbb{R}^s \to \mathbb{R}^t$ with non-zero bias. m denotes the embedding dimension, while n represents LSTM dimension. σ and $\odot$ stand for the sigmoid activation function and element-wise multiplication, respectively. $\mathbf{E} \in \mathbb{R}^{m \times L}$ is an embedding matrix. The vector $\hat{\mathbf{a}} \in \mathbb{R}^D$ holds the visual information from a particular input location of the image at time t. Thus, $\hat{\mathbf{a}}$ called context vector. Attention is a function ϕ, that computes context vector $\hat{\mathbf{a}}_t$ from the encoded vectors $\mathbf{a}_i$ (2), produced by the encoder. The attention module generates a positive number α_i for each location i on the image. This number can be interpreted as the relative importance to give to the location i, among others. Attention module is implemented as a multi-layer perceptron (MLP) with a softmax activation function, conditioned at the previous hidden state h_{t-1} (5) of the LSTM. The attention module is depicted in Fig. 1. The set of linear layers in MLP is denoted as a function f_{att}. The weights α_{ti} are computed using the following equations:

$$e_{ti} = f_{\text{att}}\,(\mathbf{a}_i, \mathbf{h}_{t-1}) \tag{6}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{p=1}^{C} \exp(e_{tp})} \tag{7}$$

The sum of weights α_{ti} (7) should be equal to 1 $\sum_{i=1}^{C} \alpha_{ti} = 1$. The context vector $\hat{a}_t$ is computed by the *attention function* ϕ with the set of encoded vectors $\mathbf{a}$ (2) and their corresponding weights α_{ti} (7) as inputs: $\hat{\mathbf{a}}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_{ti}\})$. According to the original paper function, ϕ can be either 'soft' or 'hard' attention. Due to specific task of medical image caption, function ϕ was chosen to be the 'soft' attention, as it allows model to focus more on some specific parts of X-rays from others and to detect pathologies and major organs such as heart, lung

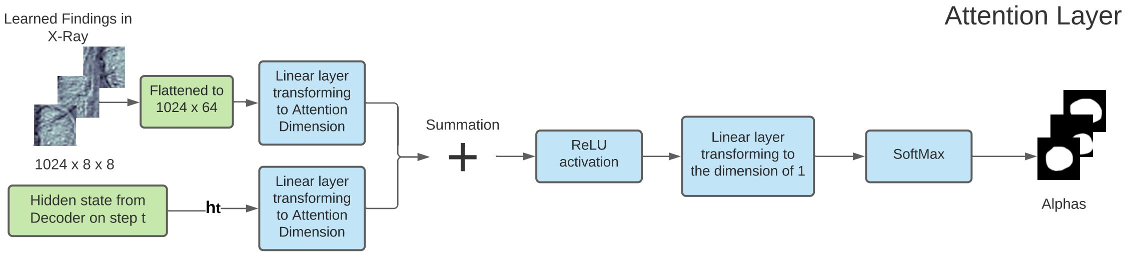

Attention Layer
Learned Findings in X-Ray
1024 x 8 x 8
Flattened to 1024 x 64
Linear layer transforming to Attention Dimension
Hidden state from Decoder on step t
ht
Linear layer transforming to Attention Dimension
Summation
+
ReLU activation
Linear layer transforming to the dimension of 1
SoftMax
Alphas

Figure 1. Attention module used in SAT.

etc. It is named as a 'deterministic soft attention' and recognized as a weighted sum : $\phi(\{\mathbf{a}_i\}, \{\alpha_{ti}\}) = \sum_i^C \alpha_i \mathbf{a}_i$. Hence, context vector can be computed as:

$$\hat{\mathbf{a}}_t = \sum_i^C \alpha_i \mathbf{a}_{ti} \tag{8}$$

The initial memory state and hidden state of the LSTM are initialized with two separate multi-layer perceptrons (init-c and init-h) with the encoded vectors $\mathbf{a}_i$ (2) for a faster convergence:

$$\mathbf{c}_0 = f_{\text{init-c}}\left(\frac{1}{C}\sum_i^C \mathbf{a}_i\right) \tag{9}$$

$$\mathbf{h}_0 = f_{\text{init-h}}\left(\frac{1}{C}\sum_i^C \mathbf{a}_i\right) \tag{10}$$

To compute the output of LSTM representing a probabilities vector the next word, a 'deep output layer'[44] was used. It looks both on the LSTM state $\mathbf{h}_t$ (5), on context vector $\hat{\mathbf{a}}_t$ (8) and the one previous word $\mathbf{z}_{t-1}$ (2):

$$P(\mathbf{z}_t|\hat{\mathbf{a}}_t, \mathbf{z}_{t-1}) = softmax(\mathbf{L}_o(\mathbf{L}_h\mathbf{h}_t + \mathbf{L}_a\hat{\mathbf{a}}_t + \mathbf{E}\mathbf{z}_{t-1})) \tag{11}$$

where $\mathbf{L}_o \in \mathbb{R}^{W\times m}$, $\mathbf{L}_h \in \mathbb{R}^{m\times n}$, $\mathbf{L}_a \in \mathbb{R}^{m\times D}$, and $\mathbf{E} \in \mathbb{R}^{m\times L}$ represent the embedding matrix.

The authors in[15] suggest to use the 'doubly stochastic attention', where $\sum_t \alpha_{ti} \approx 1$. This can be interpreted as encouraging the model to pay equal attention to every part of the image. Yet, this method is not relevant for X-rays, as each part of the chest is almost at the same position from image to image. If the model learned, e.g., that heart is in its specific position, a model does not have to search for the heart somewhere else. The model is trained in an end-to-end manner by minimizing the cross-entropy loss L_{CE} between vector with a softmaxed distribution probability of next word and true caption as $L_{CE} = -\log(P(\mathbf{z}|\mathbf{a}))$.

Generative pretrained transformer. Generative Pretrained Transformer (GPT-3)[45] is a large transformer-based language model with 1.75×10^{11} parameters, trained on 570 GB of text. GPT-3 can be used to generate realistic continuations texts from the arbitrary domain. Basically, GPT-3 is a transformer that can look at a part of the sentence and predict the next word, thus being a language model. The original transformer[46] is made up of encoder stack and decoder stack, in which encoders and decoders stacked upon each other. Whereas GPT-3 is built using just decoder blocks. One decoder block consists of Masked Self-Attention layer and Feed-Forward neural network. It is called Masked as it pays attention only to previous inputs. The input should be encoded prior to going into the decoder block. In transformers and in the GPT-3 particularly, there are two subsequent encodings: Byte Pair Token Encoding and Positional Encoding. Byte Pair Encoding (BPE) is a simple data compression technique that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. The algorithm compresses data by finding the most frequently occurring pairs of adjacent subtokens in the data and replacing all instances of the pair with a single subword. The algorithm repeats this process until no further compression is possible. Such tokenization avoids adding a special `<unk>` token to the vocabulary, as now all words can be encoded and obtained by combination of subwords from the vocabulary.

Proposed architecture

We introduce two architectures for X-ray image captioning. The overall goal of our approach is to improve the quality of Encoder-Decoder generated clinical records by using the GPT-3 language model. The suggested model consists of two parts: the Encoder, Decoder (LSTM) with an attention module and GPT-3. While Encoder with LSTM detects pathologies and indicates zones of higher attention demand, the GPT-3 takes it as input and writes a comprehensive medical report. There are two possible approaches for this task.

Approach 1 The first method consists in forcing the models to learn a joint word distribution. Within this method (Fig. 2), both models **A** and **B** output scores for the next word in a sentence. Afterwards, due to concatenating these scores and pushing them through the feed-forward neural net **C**, we get the final scores for subsequent word. Whilst the disadvantage of this approach is the following: the GPT-3 model has its own vocabulary built by the byte pair tokenizer. This vocabulary is different from the one used by the SAT model. We need to take from continuous GPT-3 distribution separate scores corresponding to the words present in the Show Attend and Tell vocabulary. This turns continuous distribution from the GPT-3 into discrete and hence, while we do not use all the potential generation power from the GPT-3.

Approach 2 The Approach 2 is shown in Fig. 3 and is based on stacked **A** and **B** models. Show Attend and Tell **A** gets an image as an input and generates a report based on the data found on X-ray with an Attention module. It learns where to focus and gives a seed for the GPT-3 **B** to continue generating text. The GPT-3 was fine-tuned on MIMIC-CXR in self-supervised manner using the Huggingface framework[47]. It learns to predict the next word in the text. The GPT-3 continues the report outputted by SAT and generates a detailed and complete clinical report based on pathologies found by SAT. Such an approach is better for the GPT-3 as it gets more context as input (from SAT) than in the first approach. Thus, the second approach performs better, and was hence chosen as the main architecture.

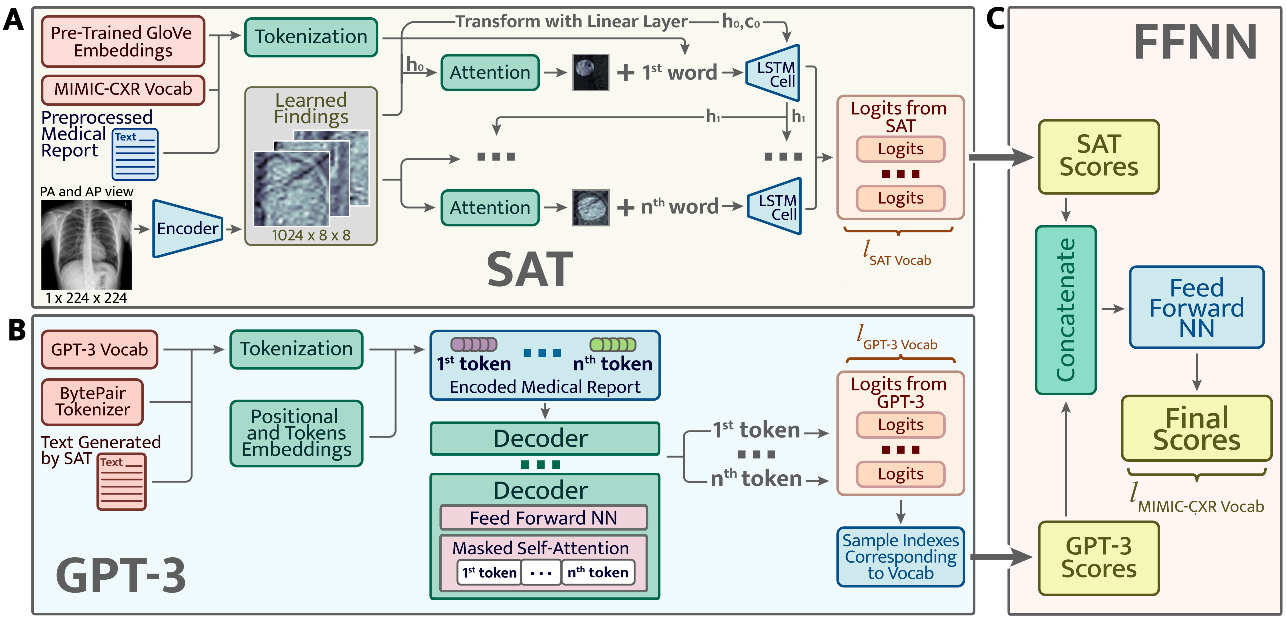


Figure 2. The first approach. Learn the joint distribution of two models.

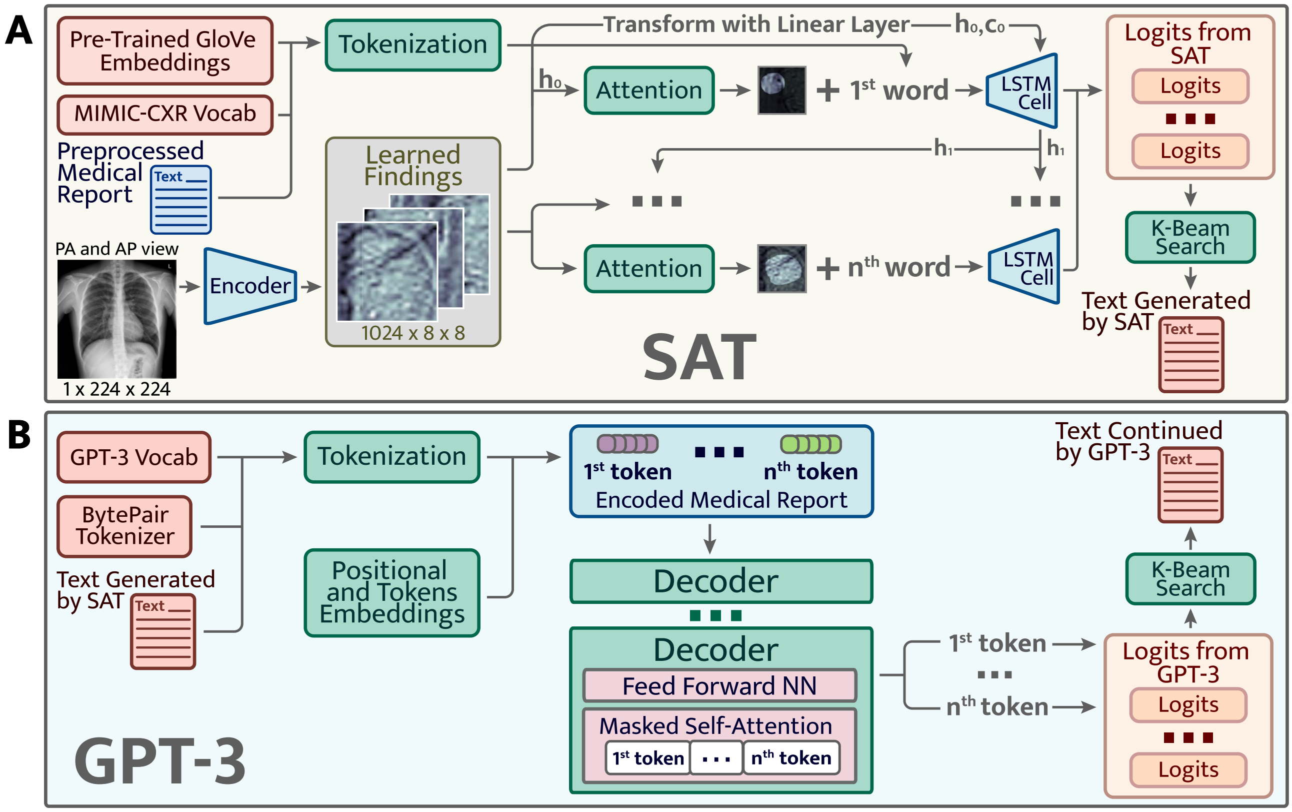


Figure 3. The second approach. Pretrained GPT-3 (**B**) continues text generated by SAT (**A**).

First language model. The first part of the suggested model is realized as the Show Attend and Tell model (SAT), the encoder to encode the image, and the LSTM for decoding into sequence. The encoder encodes the input image with 3 or 1 color channels into a smaller image with 'learned' channels. The resulted encoded images can be interpreted as a summary representation of findings in the X-ray (Eq. 2). Those encoders, pretrained on the ImageNet[48], are not suitable for the medical image caption task, as medical images do not have typical objects from the natural domain. Thus, the DenseNet-121 from[49] pretrained on the MIMIC-CXR dataset was taken. It was trained for the classification task on 18 labels : Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, Mass, Hernia, Lung Lesion, Fracture, Lung Opacity, and Enlarged Cardiomediastinum. Hence, the last classification layer was removed and features from the last convolutional layer were taken. These features were passed through the Adaptive Average Pooling layer. They can be represented by the tensor with the following dimensions: $(batchsize \times C, D, D)$ (Eq. 2). C stands for the number of channels or how many different image regions to consider. D implies the dimension of the image encoded region. Furthermore, the fine-tune method for encoder was added. It enables or disables the calculation of gradients for the encoder's parameters through the last layers. Then, at every time step, the decoder with the attention module observes the encoded small images with findings and generates a caption word by word. The Encoder output is received and flattened to dimensions $(batchsize, C, D \times D)$. Since captions are padded with a special token `<pad>`, captions are sorted by decreasing lengths and at every time-step of generating a word, an effective batch size is computed in order not to process the `<pad>` token.

The Show Attend and Tell model was trained using the Teacher-Forcing method while at each step the input to the model was the ground truth word on this step and not the previous generated word. As a result, we can consider the SAT as a language model **A**. It gets a tokenized text of length m, an image as input and outputs a vector of probabilities for the next word at each time step t:

$$\begin{aligned} \mathbf{A} : \text{text,image} \rightarrow P_1(\mathbf{z}^t|\text{true words} = \mathbf{z}^{\langle 1\rangle}\mathbf{z}^{\langle 2\rangle} \ldots \mathbf{z}^{\langle t-1\rangle}, \text{ image}), \\ t \in \{2, \ldots m, \ldots L\}, \\ P_1 \in \mathbb{R}^{m \times W_{SAT}} \end{aligned} \tag{12}$$

where W is the SAT vocabulary size and L is the length of generated report (Eq. 1). Where P_1 is computed as it is shown in the Eq. (11).

Over the training process, the LSTM outputs a word with a maximum probability after the softmax layer. Similarly to[50], we applied the K-Beam search, but only in the inference stage.

Second language model.

The second part of the architecture proposed is the GPT-3. The GPT-3 is built from decoder blocks using the transformer architecture. At the same time, the decoder block consists of masked self-attention and a feed-forward neural network (FFNN). The output yields the token probabilities, i.e., logits. The GPT-3 was pretrained separately on the MIMIC-CXR dataset and was then fine-tuned together with the SAT to enhance clinical reports.

We put a special token `<start>` at the end of the text generated by the SAT allowing the GPT-3 to understand where to start the generation process. We also used the K-Beam search after the GPT-3 generation and took the second best sentence from the output as a continuation. The pretrained GPT-3 performs as a separate language model **B** and generates good records based on the input text or tags. The GPT-3 generates report till the moment when it generates the special token `<|endoftext|>`. We denote the length of the GPT-3 generated text as l

$$\mathbf{B} : \text{text} \rightarrow P_2(\mathbf{z}^t|\text{true words} = \mathbf{z}^{<1>} \ldots \mathbf{z}^{<L>} < \mathbf{s} >), \quad t \in \{L+1, \ldots L+l\}, \tag{13}$$

Combination of two language models.

We use a combination of two models placing them sequentially: the SAT model extracts visual features from the image and allows us to focus on its specific parts. The GPT-3 provides good and comprehensive text, based on what is found by the first model. Thus, the predictions from the first model improve those of the second language model.

Evaluation metrics.

The common evaluation metrics used for image captioning are : bilingual evaluation understudy (`BLEU`)[51], recall-oriented understudy for gisting evaluation (`ROUGE`)[52], metric for evaluation of translation with explicit ordering (`METEOR`)[53], consensus-based image description evaluation (`CIDEr`)[54], and semantic propositional image caption evaluation (`SPICE`)[55]. The Microsoft Common Objects in Context[56] provides the kit with implementation of these metrics for the image caption task.

Experiments

Datasets.

For training and evaluation of medical image captioning, we use three publicly available datasets. Two of them are medical images datasets and the third one is a general-purpose one.

MIMIC-CXR The MIMIC Chest X-ray (MIMIC-CXR)[57] dataset is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports. This dataset consists of 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA.

Open-I The Indiana University Chest X-ray Collection (IU X-ray)[25] contains radiology reports associated with X-ray images. This dataset contains 7470 image-report pairs. All the reports enclose the following sections: impression, findings, tags, comparison, and indication. We use the concatenation of impression and findings as the target captions.

MSCOCO Microsoft Common Objects in Context dataset (MS COCO dataset)[58] is large-scale non-medical dataset for scene understanding. The dataset is commonly used for training and benchmark object detection, segmentation, and captioning algorithms.

Image preprocessing.

Hierarchical Data Format (HDF5)[59] dataset was used to store all images. X-rays are in gray-scale and have one channel. To process them with the pre-trained CNN DenseNet-121, we used 1 channel image. Each image was resized to the size of 224 × 224 pixels, normalized to the range from 0 to 1, and converted to the `float32` type and stored in the HDF5 dataset.

Image captions pre-processing.

Following the logic in[60], a medical report is considered as a concatenation of Impression and Findings sections, if both of these sections are empty, this report was excluded. This resulted in 360,666 DICOMs with reports for the MIMIC-CXR dataset. The text records are pre-processed by converting all tokens to lowercase, removing all non-alphanumerical tokens. For our experiments we used 75% of data for training, 24.75 % for validation and 0.25% for testing.

The MIMIC-CXR database was used to access metadata and labels derived from free-text radiology reports. These labels were extracted using the NegBio tool[21,30] that outputs one of 14 pathologies along with their severity and (or) absence. To generate more accurate reports, we added the extracted labels to the beginning of the report. This allows language models to know the summary of the report for a more precise description generation.

We additionally formed the abbreviations dictionary of 150+ words from the Unified Medical Language System (UMLS)[61]. We also extended our dictionary size with several commonly used medical terms from the Medical Concept Annotation Tool[62].

Training of the neural network.

The pipeline is implemented using PyTorch. Experiments were conducted on a server running the Ubuntu 16.04 (32 GB RAM). All models were trained with NVIDIA Tesla V100 GPU (32 GB RAM). In all experiments, we use a 5-fold cross-validation and reported the mean performance. The SAT was trained for 70 epochs with the batch size of 16, embedding dimension of 100, attention and decoder dimension of 512, dropout value 0.1. The encoder and decoder learning rates were 4×10^{-7} and 3×10^{-7}, respectively. The Cross Entropy loss was used for training. The best model is chosen according to the highest geometric mean of BLEU-n, as it is done in other works[63]. SAT was trained in Teacher-Forcing technique, while the Greedy approach is used for counting metrics. The GPT-3 small was fine-tuned with the MIMIC-CXR dataset for 30 epochs with batch size of 4, learning rate of 5×10^{-5}, the Adam epsilon of 1×10^{-8}, where the block size equals 1024, with clipping gradients, which are bigger than 1.0. It was fine-tuned in a self-supervised manner as a language model. No data augmentation was applied.

Results and discussion

Quantitative results.

The quantitative results for the baseline models, preceding works, and our models are presented in Table 1. The models were evaluated on the most common Open-I dataset, as well as on the big and rarely reported data from the MIMIC-CXR with free-text radiology reports. We implemented the most commonly used metrics for evaluation—`BLEU-n`, `CIDEr` and `ROUGE_L`. The proposed approach outperforms the existing models in terms of the NLG metrics—`BLEU-n`, `CIDEr` and `ROUGE`. `BLEU-n` measures the accuracy, `ROUGE_L` measures the recall of the generated report while `CIDEr` helps estimate the ability of the model to capture context information in the ground truth report. The higher the metrics values, the better the performance of the model.

We additionally illustrated the performance of our model in Fig. 4 containing 4 original X-ray images from the MIMIC-CXR dataset, the ground truth expert label, and the model predictions (Approaches 1 & 2). We manually underlined the similarities and identical diagnoses in texts to guide the eye. Table 2 presents the measured clinical efficacy (CE) metrics on the MIMIC-CXR dataset for the baseline models and our proposed Approaches 1 and 2. The metrics are calculated by comparing the critical radiology terminology extracted from the generated and the reference reports.

Discussion.

The first language model (SAT) learned to generate a short summary at the beginning of the report, based on the findings from a given medical image to provide the content details. This offers text generation direction seed for the second model. The preprocessing of the medical reports enabled these high metrics. We also address the biased data problem by applying domain-specific text preprocessing while using the NegBio labeller. In a radiology database, the data is unbalanced because abnormal cases are rarer than the normal ones. The NegBio labeller allowed us to get a not negative-biased diagnosis clinical records as it added short sentences at the beginning of the ground truth reports, making this task closer (in some ways) to a classification task, when the state-of-the-art models had already managed to achieve a strong performance. The SAT also provides 2D localization heatmaps of pathologies, assisting and accelerating the diagnosis process.

The second language model, the Generative Pretrained Transformer GPT-3, showed promising results in the medical domain. It successfully continued the extracted texts from the first language model, taking into consideration all the findings provided. As GPT-3 is a rather powerful transformer, it summarizes and provides more details on the findings. Natural language generation metrics suggest that using two language models subsequently provides a notable advantage. Such an approach can be considered as accurate and efficient for the medical captions generation.

One may notice a gap in the context-related performance (CIDEr)as each ground truth image is accompanied by multiple reference captions. The drawback in the CIDEr performance points to a suboptimal suitability of the generated output, whereas the Approach 2 does its best. This is due to the image-relevant n-grams occurring

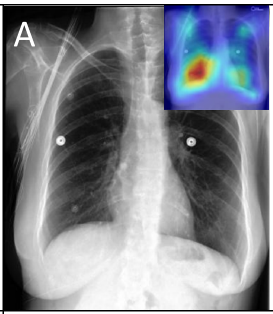
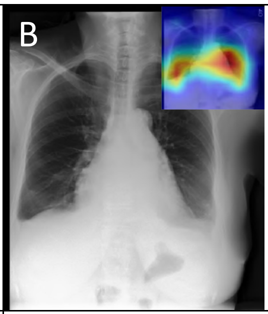
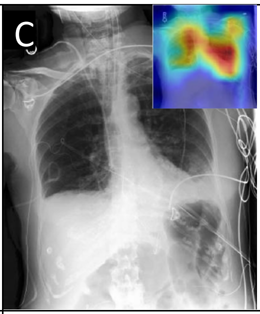
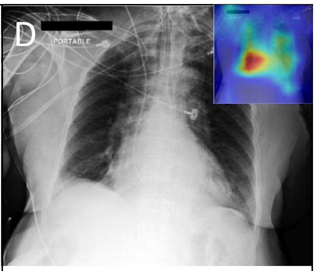
Image sample cases				
DC	No findings	Pleural effusion, Cardiomegaly, Atelectasis	Atelectasis, Pleural Effusion, Lung Opacity	Cardiomegaly, Edema, Lung Opacity
Ground truth	Lungs remain well inflated without evidence of focal airspace consolidation, pleural effusions, pulmonary edema or pneumothorax.Irregularity in the right humeral neck is related to a known healing fracture secondary to recent fall. PA and lateral views of the chest at 09:55 are submitted	1. Stable bilateral small pleural effusions andatelectasis. 2. Enlarged pulmonary artery, suggesting pulmonary hypertension. Bilateral small pleural effusions and adjacent atelectasis are overall unchanged. The heart is top-normal in size, unchanged.	There is decrease in now small right pleural effusion. There is no pneumothorax. There is a new right pacer pigtail catheter. Cardiomediastinal contours are unchanged. Lines and tubes are in standard position.	Compared to prior chest radiographs . Previous mild pulmonary edema has improved, moderate cardiomegaly and mediastinal vascular engorgement have not. ET tube, right transjugular temporary pacer lead are in standard placements and an esophageal drainage tube passes into the stomach
Approach 1	pulmonary vascularity is normal in caliber and distribution . impression : no evidence of acute pulmonary pathology with possible development of right pleural effusion .	minimal linear densities in the costophrenic angles characteristic of scarring . healed rib fractures. minimal tortuosity thoracic aorta . Multiple calcified pulmonary nodules consistent with pulmonary edema .	bilateral patchy areas of consolidation appearing possibly due to edema and/or pneumonia . anteriorchest surgery . clothing artifact appearing . anterior right upper abdomen surgery .	bilateral small pleural effusions and adjacent atelectasis are overall unchanged. The heart is top-normal in size, unchanged.
Approach 2	no findings. no pneumonia. no pleural effusion. no edema. there is little change and no evidence of acute cardiopulmonary disease. no pneumonia, vascular congestion, pleural effusion.of incidental note is an azygos fissure, of no clinical significance . this raises possibility of a normal variant.	pleural effusion present. lung opacity present. no edema. cardiomegaly present. atelectasis present. as compared to previous radiograph, there is an increase in extent of a pre existing small left pleural effusion with subsequent atelectasis at left lung bases. no new focal parenchymal opacities suggesting pneumonia.	uncertain pneumonia. pleural effusion present. lung opacity present. atelectasis present. bilateral pleural effusions, left greater than right. bibasilar opacities potentially atelectasis in setting of low lung volumes. infection be excluded.frontal and lateral views of chest demonstrate low lung volumes. there is no focal consolidation pneumothorax.	support devices present. no pneumothorax. pleural effusion present. lung opacity present. uncertain enlarged cardiomediastinum. no edema. atelectasis present. right internal jugular central line has its tip in distal superior vena cava. overall cardiac and mediastinal contours are likely stable given patient rotation on current study.

Figure 4. Image sample cases with the disease classes (DC) along with original (ground truth) and generated reports by the proposed SAT + GPT-3 model implemented as in Approach 1 and 2, respectively. Insets in the upper corners of the original images feature localization heatmaps. Heatmaps are generated using Matplotlib v.3.7.0[64].

frequently in the respective set of reference sentences. The drawback is in the sampling from the GPT-3 distribution. The Approach 2, featuring SAT followed by the GPT-3, outperformed the reported state-of-the-art (SOTA) models in all the 3 datasets considered. Notably, the proposed approach outperforms SOTA models on MIMIC-CXR, demonstrating the highest performance in all the metrics. The performance for the main evaluation dataset, the MIMIC-CXR, is measured by the CE metrics using micro-averaging and demonstrates 0.861 for the proposed SAT + GPT-3 Approach 2 model vs. 0.840 with the Approach 1, and 0.743 for the SAT, respectively, as reported in Table 2.

Examples of the reports generated jointly via the SAT + GPT-3 with Approaches 1 and 2 are shown in Fig. 4. One may notice that some generated sentences coinside with the ground truth. For example, in both generated and the true reports, for the first X-ray it reads "no acute cardiopulmonary abnormality". Some sentences are close in their meaning, even if they are different in terms of chosen words and n-grams ("no pneumonia. no pleural effusion. no edema. ..." compared to " without pulmonary edema or pneumothorax").

Conclusions

We introduced a new technique of combining two language models for the medical image captioning task. Principally, the new preprocessing and squeezing approaches for clinical records were implemented along with a combined language model, where the first component is based on attention mechanism and the second one represents a generative pretrained transformer. The proposed combination of the models generates a descriptive textual summary with essential information on found pathologies along with their location and severity. Besides, the 2D Grad-CAM[67] heatmaps localize each pathology on the original scans. The results, measured with the natural language generation metrics on both the MIMIC-CXR and the Open-I datasets, speak for an efficient applicability to the chest X-ray image captioning task. This approach also provides well-interpretable results and allows to support clinical decision making.

We investigated various approaches to automatic generation of X-ray image captioning. We proved that the SAT is a strong baseline, outperforming models with Transformer-based decoders. With the help of GPT-3 pre-trained language model, we managed to improve this baseline. The simple method, where the GPT-3 model

	Model	CIDEr	ROUGE_L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MIMIC-CXR	S &T[12]	0.886	0.300	0.307	0.201	0.137	0.093
	Original SAT[15]	0.967	0.288	0.318	0.205	0.137	0.093
	TieNet[19]	1.004	0.296	0.332	0.212	0.142	0.095
	NLG[29]	1.153	0.307	0.352	0.223	0.153	0.104
	SAT†	1.986	0.478	0.634	0.549	0.451	0.383
	Approach 1	1.974	0.477	0.622	0.573	0.497	0.401
	Approach 2	**1.989**	**0.480**	**0.725**	**0.626**	**0.505**	**0.418**
Open-I	Co-Attention[60]	0.327	0.447	0.517	0.386	**0.306**	**0.247**
	TieNet[19]	–	0.311	0.330	0.194	0.124	0.081
	CNN-RNN[12]	0.111	0.267	0.316	0.211	0.140	0.095
	LRCN[65]	0.190	0.278	0.369	0.229	0.149	0.138
	ATT-RK[18]	0.155	0.323	0.369	0.226	0.151	0.108
	CDGPT2[37]	0.257	0.289	0.387	0.245	0.166	0.111
	Original SAT[15]	0.320	0.361	0.433	0.281	0.194	0.138
	SAT†	0.699	0.413	0.407	0.258	0.210	0.125
	Approach 1	0.687	0.402	0.450	0.299	0.224	0.141
	Approach 2	**0.701**	**0.450**	**0.520**	**0.390**	0.296	0.235
MS-COCO	BRNN[66]	–	–	0.642	0.451	0.304	0.203
	Original SAT[15]	–	–	0.718	0.504	0.357	0.250
	SAT†	1.300	0.592	0.815	0.663	0.516	0.395
	Approach 1	1.298	0.505	0.818	0.664	0.509	0.385
	Approach 2	**1.360**	**0.606**	**0.821**	**0.672**	**0.529**	**0.409**

Table 1. Reported mean performance using word-overlap metrics for two medical radiology datasets and one non-medical for general purpose. Models labelled with † stand for the models we implemented and trained with the preprocessed MIMIC-CXR data. Other results are cited from the original papers. BLUE-n denotes the BLEU score that uses up to n-grams. The best performance in each configuration is in bold.

Model	Accuracy	Precision	Recall	F1-Score
S &T[12]	0.423	0.084	0.066	0.072
Original SAT[15]	0.703	0.181	0.134	0.144
TieNet[19]†	0.741	0.265	0.178	0.197
NLG[29]†	0.792	0.413	0.286	0.317
SAT†	0.743	0.166	0.121	0.129
Approach 1	0.840	0.420	0.303	0.134
Approach 2	**0.861**	**0.445**	**0.351**	**0.369**

Table 2. The clinical efficacy (CE) metrics on the MIMIC-CXR dataset. The best results are highlighted in bold. Models labelled with † stand for the models we implemented and trained with the preprocessed MIMIC-CXR data. Other results are cited from the original papers. .

finishes the report extracted by the Show-Attend-Tell model, yields significant improvements to the standard text generation scores. Recent advancements in interactive training, such as active learning[68] and dialog-based ChatGPT[69], have the potential to improve the performance of medical image captioning models even further. This is an area of research that will be explored in the future.

Data availability

All data generated or analysed during this study are included in this published article. The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 24 October 2022; Accepted: 8 March 2023
Published online: 13 March 2023

References

1. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
2. Demner-Fushman, D. *et al.* Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310 (2016).

3. Chan, Y.-H. *et al.* Effective pneumothorax detection for chest X-ray images using local binary pattern and support vector machine. *J. Healthc. Eng.* **2018**, 1–11 (2018).
4. Maghdid, H. S. *et al.* Diagnosing covid-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. In *Multimodal Image Exploitation and Learning 2021* Vol. 11734, 117340E (International Society for Optics and Photonics, 2021).
5. Monshi, M. M. A., Poon, J. & Chung, V. Deep learning in generating radiology reports: A survey. *Artif. Intell. Med.* **106**, 101878 (2020).
6. García Gilabert, J. Image Captioning using pre-trained GPT-2 models. Ph.D. thesis, Universitat Politècnica de València (2022).
7. Chen, J., Guo, H., Yi, K., Li, B. & Elhoseiny, M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18030–18040 (2022).
8. Sermesant, M., Delingette, H., Cochet, H., Jaïs, P. & Ayache, N. Applications of artificial intelligence in cardiovascular imaging. *Nat. Rev. Cardiol.* **18**, 600–609 (2021).
9. Gurgitano, M. *et al.* Interventional radiology ex-machina: Impact of artificial intelligence on practice. *La radiologia medica* **126**, 998–1006 (2021).
10. Belikova, K., Rogov, O. Y., Rybakov, A., Maslov, M. V. & Dylov, D. V. Deep negative volume segmentation. Sci. Rep. **11** (2021).
11. Pavlopoulos, J., Kougia, V. & Androutsopoulos, I. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, 26–36 (Association for Computational Linguistics, 2019).
12. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164 (2015).
13. Shin, H.-C. *et al.* Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
14. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)* (2016).
15. Xu, K. *et al.* Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning* Vol. 37, ICML'15, 2048-2057 (JMLR.org, 2015).
16. Donahue, J. *et al.* Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 677–691 (2017).
17. Zhang, Z., Xie, Y., Xing, F., McGough, M. & Yang, L. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3549–3557 (2017).
18. You, Q., Jin, H., Wang, Z., Fang, C. & Luo, J. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4651–4659 (2016).
19. Wang, X., Peng, Y., Lu, L., Lu, Z. & Summers, R. M. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
20. Jing, B., Xie, P. & Xing, E. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* Vol. 1: Long Papers, 2577–2586 (Association for Computational Linguistics, 2018).
21. Wang, X. *et al.* Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471 (2017).
22. Rodin, I., Fedulova, I., Shelmanov, A. & Dylov, D. V. Multitask and multimodal neural network model for interpretable analysis of x-ray images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2019).
23. Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L. J. & Bradley, A. P. Producing radiologist-quality reports for interpretable deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (IEEE, 2019).
24. Yuan, J., Liao, H., Luo, R. & Luo, J. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* (eds Shen, D. *et al.*) 721–729 (Springer International Publishing, 2019).
25. Demner-Fushman, D. *et al.* Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310 (2016).
26. Zhang, Y. *et al.* When radiology report generation meets knowledge graph. *Proc. AAAI Conf. Artif. Intell.* **34**, 12910–12917 (2020).
27. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225 (2017).
28. Shvetsova, N., Bakker, B., Fedulova, I., Schulz, H. & Dylov, D. V. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access* **9**, 118571–118583 (2021).
29. Liu, G. *et al.* Clinically accurate chest X-ray report generation. In *Proceedings of the 4th Machine Learning for Healthcare, Proceedings of Machine Learning Research* Vol. 106 (eds Doshi-Velez, F. *et al.*) 249–269 (PMLR, 2019).
30. Peng, Y. *et al.* Negbio: A high-performance tool for negation and uncertainty detection in radiology reports. In *AMIA Summits on Translational Science Proceedings* Vol. 2017 (2017).
31. Ni, J., Hsu, C.-N., Gentili, A. & McAuley, J. Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1954–1960 (Association for Computational Linguistics, Online, 2020).
32. Syeda-Mahmood, T. *et al.* Chest X-ray report generation through fine-grained label learning. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*, 561–571 (Springer International Publishing, 2020).
33. Liu, J. *et al.* Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).
34. Cohen, J. P., Hashir, M., Brooks, R. & Bertrand, H. On the limits of cross-domain generalization in automated X-ray prediction. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research* (eds Arbel, T. *et al.*) 136–155 (PMLR, 2020).
35. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long and Short Papers), 4171–4186 (Association for Computational Linguistics, 2019).
36. Ziegler, Z. M., Melas-Kyriazi, L., Gehrmann, S. & Rush, A. M. Encoder-agnostic adaptation for conditional language generation. arXiv:1908.06938 (2019).
37. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M. & Fahmy, A. Automated radiology report generation using conditioned transformers. *Inform. Med. Unlocked* **24**, 100557 (2021).
38. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449 (Association for Computational Linguistics, Online, 2020).
39. Xiong, Y., Du, B. & Yan, P. Reinforced transformer for medical image captioning. In *Machine Learning in Medical Imaging*, 673–680 (Springer International Publishing, 2019).
40. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269 (2017).

41. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015* (eds Bengio, Y. & LeCun, Y.) (Conference Track Proceedings, 2015).
42. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
44. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
45. Brown, T. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. *et al.*) 1877–1901 (Curran Associates Inc, 2020).
46. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. *et al.*) (Curran Associates Inc, 2017).
47. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).
48. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).
49. Cohen, J. P. *et al.* Torchxrayvision: A library of chest X-ray datasets and models. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research* Vol. 172 (eds Konukoglu, E. *et al.*) 231–249 (PMLR, 2022).
50. Wiseman, S. & Rush, A. M. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1296–1306 (Association for Computational Linguistics, 2016).
51. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL'02* (Association for Computational Linguistics, 2001).
52. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81 (Association for Computational Linguistics, 2004).
53. Banerjee, S. & Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72 (Association for Computational Linguistics, 2005).
54. Vedantam, R., Zitnick, C. L. & Parikh, D. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575 (2015).
55. Anderson, P., Fernando, B., Johnson, M. & Gould, S. SPICE: Semantic propositional image caption evaluation. In *Computer Vision—ECCV 2016*, 382–398 (Springer International Publishing, 2016).
56. Fang, H. *et al.* From captions to visual concepts and back. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015).
57. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6** (2019).
58. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014*, 740–755 (Springer International Publishing, 2014).
59. Koziol, Q. *et al.* HDF5. In *Encyclopedia of Parallel Computing*, 827–833 (Springer US, 2011).
60. Jing, B., Xie, P. & Xing, E. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* Vol. 1: Long Papers (Association for Computational Linguistics, 2018).
61. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **32**, 267D – 270 (2004).
62. Kraljevic, Z. *et al.* Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.* **117**, 102083 (2021).
63. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, 311–318 (Association for Computational Linguistics, 2002). https://doi.org/10.3115/1073083.1073135.
64. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95. https://matplotlib.org/stable/index.html (2007).
65. Donahue, J. *et al.* Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 677–691 (2017).
66. Karpathy, A. & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015).
67. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626 (2017).
68. Shelmanov, A. *et al.* Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Association for Computational Linguistics, 2021).
69. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).

Acknowledgements

The authors of this paper thank Alexander Panchenko and Alexander Shvets for the helpful discussion. The work was supported by Philips in 2021.

Author contributions

All authors participated in data analysis, manuscript writing, and final text review. A.S. and O.R. contributed equally to the network architecture design.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.V.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.