

## Article

# Lenke Classification Report Generation Method for Scoliosis Based on Spatial and Context Dual Attention

Yu Tang, Zhiqin He \*, Qinmu Wu, Xiao Wang  and Yuhang Wang

College of Electrical Engineering, Guizhou University, Guiyang 550025, China; ttyyqqyy@163.com (Y.T.); qmwu@gzu.edu.cn (Q.W.); xwang9@gzu.edu.cn (X.W.); 15209286372@163.com (Y.W.)

\* Correspondence: zqhe@gzu.edu.cn

**Abstract:** The scoliosis report is a diagnosis made by the clinician looking at X-ray images of the spine. However, with numerous images, writing the report can be time-consuming and error-prone. Therefore, this paper proposes an automatic generation model of the end-to-end scoliosis Lenke classification report. The model automatically generates a short diagnostic text to explain the results of the classifiers' Lenke classification diagnosis of scoliosis. Instead of reproducing the original diagnostic report, the original diagnostic report is described as a short sentence with diagnostic significance. In the model, the CBAM attention module is added to the residual's path of ResNet-50 to extract key regional features of the image, and the improved Long Term and Short Term Memory Network (M-LSTM) fusion attention mechanism with additional gated operations is used as the decoder to generate more relevant description statements. The model was verified on the scoliosis dataset from Guizhou Orthopaedic Hospital, and the generated diagnostic text obtained good scores on BLEU and CIDEr evaluation indexes, and also satisfactory scores on the evaluation criteria of five professional clinicians. Therefore, the diagnostic text generated by this method had good performance in accuracy and semantic expression.

**Keywords:** Lenke typing report generation; dual attention mechanisms; spatial attention; context attention



**Citation:** Tang, Y.; He, Z.; Wu, Q.; Wang, X.; Wang, Y. Lenke Classification Report Generation Method for Scoliosis Based on Spatial and Context Dual Attention. *Appl. Sci.* **2023**, *13*, 7981. <https://doi.org/10.3390/app13137981>

Academic Editor: Jan Egger

Received: 19 May 2023

Revised: 30 June 2023

Accepted: 6 July 2023

Published: 7 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, the generation method of image description [1], based on deep learning, has made great progress in the field of natural images, but in the field of medical images there is still a lack of effective methods that can automatically analyze diseases in medical images and generate diagnostic text. The reason is that the automatically generated disease diagnosis text should not only conform to the grammatical rules of natural language, that is, the formal cohesion should be good, but should also ensure semantic coherence. Therefore, understanding and describing complex medical radiology images in accurate natural language and automatically generating diagnostic reports remains challenging [2].

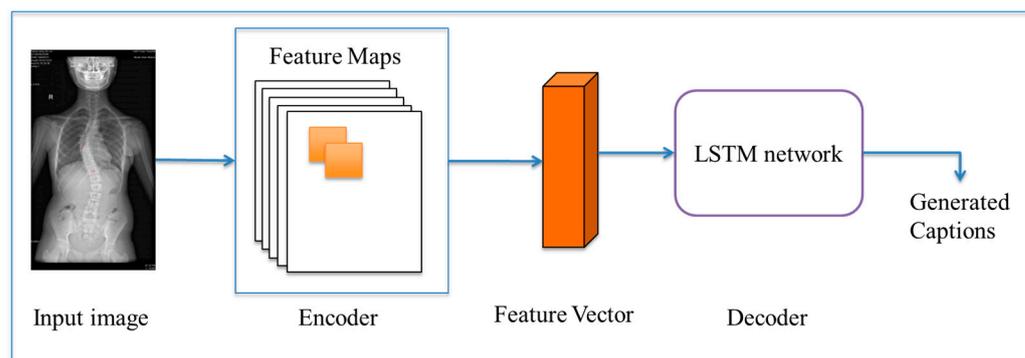
At present, medical image caption models based on deep learning [3] are all obtained by fine-tuning based on an encoder–decoder structure or by using models pre-trained on natural images [4] to train public medical image datasets. Although the encoder–decoder structure has been continuously improved to achieve good performance in the description of medical images, the diagnostic reports generated by them still contain numerous repeated words [5], the generated descriptive sentences are incoherent, and the disease information described is inconsistent with the original disease. In order to solve this problem, the idea of attention mechanisms was used as a reference, and attention mechanisms were added to the traditional codec structure. According to the hidden state of the RNN, the attention mechanism dynamically selects the local feature vector of the image related to the generated word to guide the generation of the word at the current moment. The method based on attention mechanisms can make full use of image feature information and improve the effect of description generation [6].

However, the traditional medical image description generation method based on attention mechanisms only integrates the word and visual information of the current moment when predicting the target word, ignoring the role of context information in generating the target word, resulting in no coherence and differences between the generated description and the reference description. Therefore, this shows that when generating target words, the model should not only focus on the current visual information but also the previous visual information to ensure visual coherence and conform to human habits. In addition, in traditional medical image description methods, they try to generate complete and complex diagnostic texts based on radiological reports [7]. However, in terms of radiological images, attempts to copy highly uniform radiological diagnostic reports will lead to inconsistent diagnostic descriptions in the output, which cannot be used as a true explanation of diseases in the images. Even in the most cutting-edge work in the field [7], the resulting text is nowhere near as good as that of a professional physician and cannot replace a physician in the automated diagnosis of images.

Aiming at the problems in the traditional methods of image description generation above, this paper proposes a method of generating a scoliosis Lenke classification report based on the dual attention of space and context. Instead of trying to reproduce the entire scoliosis Lenke classification and diagnosis report [8], we only generated the most important sentences. We tried to train a model that generated a simplified but meaningful description as a diagnostic report that could be certified by a clinician and can be used as a reasonable justification for a physician's decision. This method changes the way that the original diagnosis report is used as the training label in the traditional medical image description generation model. Based on the visual features of the spine images, they were innovatively hand-marked into a new set of short sentences as a label for model training. The corresponding label sentences for each image had the same sentence structure, for example: "This spine, with or without curvature in the proximal thoracic, with or without curvature in the main thoracic, with or without curvature in the lumbar regions, belongs to the Lenke-N type" or "asymptomatic cases". In addition, the traditional coding-decoding structure has been improved [9–11]. In the residual path of ResNet50 [12], a convolution block attention mechanism focusing on effective features in the space and channels was added to serve as a feature encoder to deepen the extraction of key features of the spine. Using the improved long short-term memory network M-LSTM [13] as a decoder, additional gating operations were introduced in front of LSTM [14], so that the input and state of the network were computed in multiple rounds of interaction before the internal calculation of the network, and the context attention mechanism was integrated so that the language modeling ability of the interactive representation between the input and the context became richer. Finally, the model was verified on the proprietary spinal X-ray dataset from Guizhou Orthopaedic Hospital, and the generated diagnostic description obtained satisfactory scores in BLUE, CIDEr and other indicators. The generated diagnostic text was compared with the original diagnostic report, including all key descriptions in the original diagnostic report. This shows that our model can automatically generate simple, high-quality Lenke classification reports for scoliosis.

## 2. Related Work

Medical image description generation has great practical significance, which attracts many researchers to carry out targeted exploration. At present, medical image description models based on deep learning methods are all based on an encoder-decoder structure (as shown in Figure 1), and deep CNN and RNN are utilized in the encoder-decoder structure of these models [15]. In fact, CNN extracts significant features from the input medical images in the form of fixed-length vectors. RNN or its variant networks, such as long short-term memory (LSTM) and gated reentry unit (GRU), are used as language models [16] to translate fixed-length vectors obtained from CNN into semantically and syntactically correct disease diagnosis texts.



**Figure 1.** Encoder–decoder-based caption generation.

Shin et al. [17] applied the coding–decoder structure for the first time to the task of generating medical diagnostic reports for chest X-rays, and the sentences generated by this method were mainly descriptions of disease classification and detection. At the coding end, they selected network-network [18] and Google network [19] CNN to extract image features and classify them, and obtained 17 disease labels corresponding to MESH statements. These terms were the most commonly used sentences to describe chest diseases without obtaining repeated results from other MESH terms. Although Shin et al. [17] obtained a better performance on Google, only short sentences of five words were generated to describe classified diseases, rather than a fluent and coherent report. Wu et al. [20] modified the basic framework of Vinyals et al. [16] to generate a description of retinopathy. They used a pre-trained CNN as an encoder, using exit and integration techniques to encode 37,000 fundus images, resulting in feature coding vectors that were fed into the LSTM and decoded into a caption. The method of Wu et al. [20] only generated the description of abnormal features in the images, and the description of disease information was not detailed enough.

The subtitle generation model of ultrasound images by Zeng et al. [21] was divided into two modules, an encoder module and a language generation module, and the two modules needed to be trained separately. They used a region detection model, Faster-RCNN [22], as an encoder to extract disease features and were able to locate diseased areas in ultrasound medical images. Finally, the language generation module generated the description of the critical area of the lesion and could generate the prediction box to locate the lesion area. Compared with other models, this model had the advantages of a short training time and fewer parameters, but the disadvantage was that the prediction error rate of language and words was high in the generated description statements. The decoding phase of the above methods could not obtain enough useful information, and finally the decoding of the sentence could not reach its ideal effect.

Therefore, many researchers have begun to add attention mechanisms to the coding–decoding structure. For example, Zhang et al. [23] proposed the MDNet model, which innovatively added attention mechanisms to the description task of medical images. The attention module they used was able to increase the model’s ability to extract features, successfully enabling direct mapping from bladder cancer images to diagnostic reports. They used ResNet as an encoder to extract image features and extended it to skip connections to solve the problem of disappearing gradients during training. The image description model they proposed achieved a relatively advanced performance, but the generated diagnostic report only contained the description words of five bladder features, not a complete bladder radiology film diagnostic report. Jing et al. [24] proposed an image description model with a co-attention mechanism that divided images into regions of the same size, extracted visual features from the last convolutional layer of VGG-19 [25], and calculated the weighted sum of semantic features and visual features, respectively. The visual context vector and semantic context vector were obtained, respectively, and then the final feature vector was obtained by concatenation. The decoding part adopted the hierarchical structure LSTM, composed of sentence LSTM and word LSTM. While their hierarchical LSTM achieved

good performance, the reports they generated contained repeated words, and the absence of contextual consistency in their hierarchical model could be the cause of the repetition. To further address this problem, Wang et al. [7] proposed an approach similar to Jing et al. [24], but they encoded input images using ResNet-50 and generated image titles using RNN-based LSTM, but their LSTM was flat, instead of the layered one used in [24]. However, from their experiments on the chest X-ray dataset, the descriptive results were worse than the OpenI results of Jing et al. [24], which may be because the decoding effect of flat LSTM was not as good as that of layered LSTM.

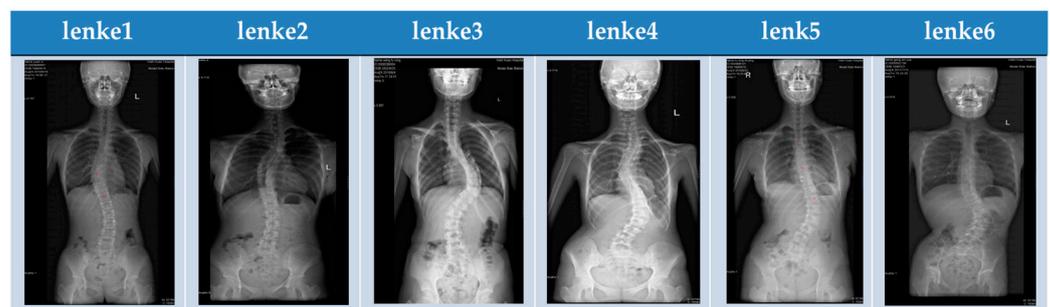
### 3. Materials and Methods

#### 3.1. Datasets

We used 13,200 positive X-ray images of scoliosis from patients with scoliosis in Guizhou Orthopedics Hospital, including 2820 Lenke-1, 1820 Lenke-2, 980 Lenke-3, 1260 Lenke-4, 2820 Lenke-5, and 2500 Lenke-6, 1000 asymptomatic sheets, and corresponding diagnostic reports. The diagnostic report was then used for training after statement processing. Specifically, we used 60% of the dataset for training (7920 images), 20% for validation (2140 images), and 20% for testing (2140 images). Details of the dataset are shown in Table 1, and examples of spinal X-ray images are shown in Figure 2.

**Table 1.** Experimental dataset details.

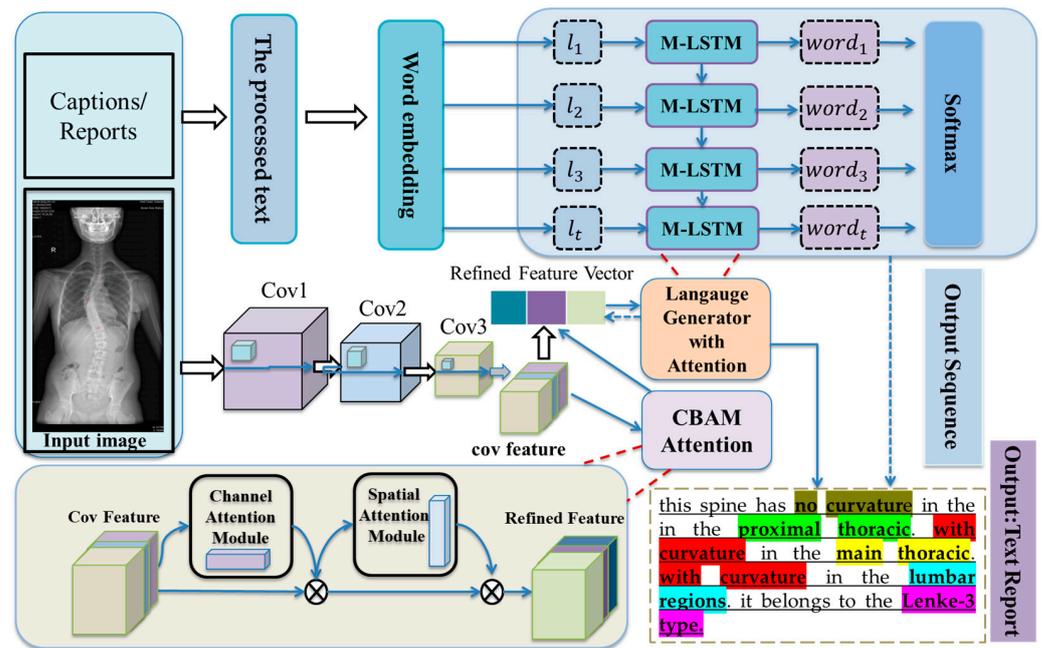
Dataset	Total	Lenke-1	Lenke-2	Lenke-3	Lenke-4	Lenke-5	Lenke-6
train	7920	1800	1320	560	840	1800	1600
test	2140	520	300	200	200	520	400
val	2140	500	200	220	220	500	500



**Figure 2.** Lenke classification radiogram of scoliosis.

#### 3.2. Overall Framework

This paper proposes an automatic generation model for end-to-end Lenke classification reports for scoliosis. The model uses ResNet-50 [11], which integrates the CBAM channel and spatial attention mechanism, as the feature encoder. Image features extracted by the ResNet50 network are re-calibrated into maps and weighted for different feature channels after compression, excitation, and spatial attention operations in the CBAM module. Therefore, the network is able to extract better-correlated features from spinal images for encoding. The improved long and short-term memory network M-LSTM fusion attention mechanism is used as the decoder. After the input of the coding vector, the features requiring active attention in the corresponding image region are dynamically selected according to the corresponding feature weights, and the image region with a high correlation with the current word is focused on. The key parts of the image are selectively focused by selecting a subset of the feature vector, and more reasonable description words are generated. The overall framework of the model is shown in Figure 3.



**Figure 3.** Overall illustration of our module. Image features are extracted by feature extraction model, processed by CBAM attention module, and then transmitted to M-LSTM combined with attention focusing image key parts, and finally prediction results are outputted.

The input to the model is image  $I$ , and the sentence sequence corresponding to  $I:L^T = \{l_1, l_2, l_3 \dots l_t\}$ . First, image input  $I$  is convolved by ResNet50 feature encoder to extract image feature  $f_i$ . Then, the CBAM module added in the residual path is used for compression, excitation and spatial attention operation, and different weight  $\alpha_i$  is assigned to the characteristic channels one by one, Finally, the output  $f'_i = f_i \times \alpha_i$  is obtained by operation Scale, and  $f'_i$  is input to the M-LSTM decoder as a new feature encoding vector. The mapping process is  $h_{-1} = w_f f'_i + b_f$ .  $w_f$  is the weight parameter that ResNet50 network needs to train, and  $b_f$  is the offset value.  $h_{-1}$  is the initial value input in the M-LSTM decoder.  $L^T$  is another input word vector of the model. It is encoded by the label word  $S$  of the training image through one hot encoding mode to obtain a word encoding vector  $m$  with a fixed length of 15 and multiplied by the word embedding matrix  $w_{emf}$  to obtain:

$$L^T = w_{emf} \times m \tag{1}$$

$$h_t = ExtLSTM(h_{t-1}, L_T) \tag{2}$$

$$p_t = Softmax(w_c \times h_t + b_c) \tag{3}$$

In Equations (2) and (3),  $h_t$  is the final output value of M-LSTM.  $w_c$  and  $b_c$  are the weight parameter and bias parameter of Softmax classifier, respectively, and  $p_t$  is the output of image description statement of Softmax classifier.

In the training process of the model, input image is  $I$ , output  $S = \{s_1, s_2, s_3 \dots s_t\}$ , which corresponds to the description sentence  $I$ , and the probability of describing output  $S$  is Equation (4). The logarithm of its likelihood function is taken to obtain logarithmic likelihood function Equation (5):

$$P(S | I; \theta) = \prod_{t=0}^N P(s_t | s_0, s_1, \dots, s_{t-1}, I; \theta) \tag{4}$$

$$\log P(S | I; \theta) = \sum_{t=0}^N \log P(s_t | s_0, s_1, \dots, s_{t-1}, I; \theta) \tag{5}$$

The training objective of the model is to maximize the logarithmic likelihood sum of all training samples, where  $\theta$  is the parameter that the model needs to be trained, and the solution of  $\theta$  can be converted into the solution of an optimal problem:

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{(I,S)} \log p(S | I; \theta) + \lambda \theta \tag{6}$$

where the first term is the model loss function, and the second term is the regularization loss. In the training process, changes and optimization are constantly adopted to maximize the probability of image description corresponding to the input image data.

### 3.3. Feature Extraction

ResNet solves the degradation problem of deep networks through residual learning and can extract image features better. ResNet directly connects and combines the output of convolutional layers with different depths, which deepens the convolutional neural network and controls the complexity of the network at the same time. It also adds a residual unit through a short-circuit mechanism, which opens a channel next to the network so that the input can reach the output directly. The object of optimization is changed from the original fitting output  $F(x)$  to the sum of output and residual path input  $x$ :  $H(x) = F(x) + x$ . Such a structure prevents the residual network from losing the feature information extracted from the previous layer due to too much depth in the process of model training. Therefore, with the deepening of the ResNet network, its learning ability will become stronger, and its ability to extract image features will be greatly improved compared with the traditional convolutional neural network. In addition, ResNet’s downsampling channel uses convolution, which changes the original fully connected layer. ResNet network models are Bottlenecks filled with residual modules, but ResNET-50, ResNET-101, and Resnet-152 differ only in the number of Stage4 Bottleneck2 modules.

In this paper, ResNet-50 [12] was selected for image feature extraction. While maintaining the depth and complexity of the network, ResNet-50 avoids the problem of the traditional convolutional neural network possibly causing gradient disappearance and degradation due to excessive depth, and it has faster convergence speed and is easier to learn. Figure 4 shows the network structure of ResNet-50.

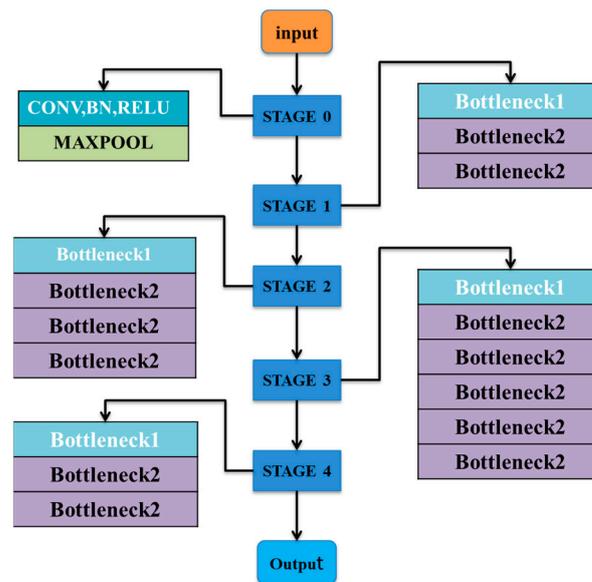


Figure 4. The bottleneck structure of ResNet 50.

### 3.4. CBAM Attention

Although Resnet50 extracts multiscale features to ensure the diversity of features, it cannot focus on the importance of features. Adding the SE module after the Resnet50

module can allow a better focus on the effective characteristics of sampling. CBAM is an attentional mechanism module combining space and channel. Compared with SENet [26], CBAM achieves better performance in image feature extraction. In deep learning, due to the difference in weight value between each channel, the product of each weight value and the value of the original channel is the characteristic value really required by deep learning. Based on SENet, the CBAM module considers the importance of pixels in different positions of each channel, obtains the importance of feature maps in channel and space by learning, and changes the weight value of different features to extract a more effective feature map. Suppose the input feature maps of the module in the middle of the network are  $F \in R^{C \times H \times W}$ ,  $C, H$ , and  $W$  are the number, height and width of input feature graph channels, respectively, and the output of feature graph  $F'$  is obtained through the operation of the one-dimensional channel attention mechanism  $M_C = R^{C \times 1 \times 1}$ , as shown in Equation (7). After the two-dimensional space attention mechanism  $M_s = R^{1 \times H \times W}$ , the feature graph output  $F''$  is obtained, as shown in Equation (8), where  $\otimes$  is the feature graph multiplied element by element.

$$F' = M_C(F) \otimes F \tag{7}$$

$$F'' = M_s(F') \otimes F' \tag{8}$$

In Formula (7), the channel attention pays attention to the meaningful learning features in the feature graph. By using maximum pooling and average pooling, the feature graph is compressed in the spatial dimension to obtain two different spatial background description factors,  $F_{max}^C$  and  $F_{avg}^C$ . The shared network of multi-layer sensing units is used to calculate the background description factors to obtain the channel attention feature graph, as shown in Formulas (9) and (10), with  $W_0 \in R^{C/r \times c}$ ,  $b$ , Relu as the activation function after  $W_0$ .

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{9}$$

$$M_c(F) = \sigma\left(W_1\left(W_0\left(F_{avg}^C\right)\right) + W_1\left(W_0\left(F_{max}^C\right)\right)\right) \tag{10}$$

In Formulas (10) and (11), the spatial attention pays more attention to the location information in the feature graph. Maximum pooling and average pooling are adopted to obtain two different feature description factors,  $F_{max}^s \in R_{1 \times H \times W}$  and  $F_{avg}^s \in R_{1 \times H \times W}$ , in the channel dimension. The calculation process is as shown in Formulas (12) and (13). Finally, the two features are combined and the convolution operation is used to obtain the spatial attention feature figure  $M_s(F) \in R_{H \times W}$ .

### 3.5. Decoder M-LSTM

Long-term and short-term memory networks can effectively improve the long-term dependence problem common in recurrent neural networks, but the network model still has some problems, such as weak generalization ability, weak language modeling ability, and poor performance in the face of context decorrelation. By introducing additional gating operations, the MOGRIFIER-LSTM [13] model enables multiple rounds of interactive calculation between input and hidden state before input to LSTM, which enables richer interactive representation between input and context and enhances the context modeling ability of the model. The specific process of interactive calculation between input and hidden states is shown in Figure 5.

As can be seen from the figure above, multiple rounds of interactive calculation were carried out between input  $x$  and hidden state  $h$  before input to LSTM, where  $x^{-1}$  is input  $x$  and  $h^0$  is hidden state  $h_{prev}$ . The order of interactive calculation is  $r$ , which is a hyperparameter. When  $r = 0$ , the network does not carry out additional gating operation, and the calculation method is as follows:

$$x^i = 2\sigma\left(Q^i h_{prev}^{i-1}\right) \odot x^{i-2}, \text{ for odd } i \in [1, \dots, r] \tag{11}$$

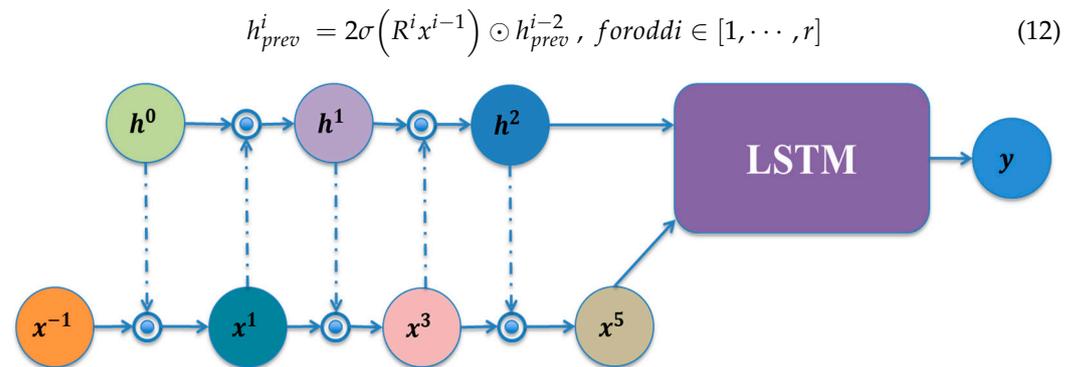


Figure 5. Interactive calculation process diagram.

The input and hidden state obtained through multiple rounds of iterative interaction fusion are inputted into LSTM. The calculation method is as follows:

$$f = \sigma(W^{fx}x + W^{fh}h_{prev} + b^f) \tag{13}$$

$$i = \sigma(W^{ix}x + W^{ih}h_{prev} + b^i) \tag{14}$$

$$h = o \odot \tanh(c) \tag{15}$$

$$c = f \odot c_{prev} + i \odot j \tag{16}$$

$$h = o \odot \tanh(c) \tag{17}$$

where  $\sigma$  is sigmoid activation function,  $w$  is weight matrix, and  $b$  is bias.

### 3.6. M-LSTM with Contextual Attention

The image description model with the introduction of the attention mechanism will generate each word of the description statement according to the significant part of the image that is currently concerned, so that the decoder can selectively focus on some parts of the image by selecting a subset of feature vectors and pay attention to different image regions at different moments of decoding to generate more reasonable words. What is calculated by the attention mechanism is the context vector that captures the visual information of a particular area. The calculation method is as follows:

$$e_{ti} = f_{att}(a_i, h_{t-1}) \tag{18}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^L \exp(e_{tk})}$$

where  $\alpha_i$  is the region to be calculated,  $h_{t-1}$  is the information of the previous time  $t - 1$ ,  $f_{att}$  is the multi-layer perceptron scoring function of the two pieces of information of the coupled calculation area  $i$  and time  $t$ , and  $\alpha_{ti}$  is the weight, which measures the degree of attention of the image area  $i$  at moment  $t$ .

$$p(s_{t,i} = 1 | s_{j < t}, a) = \alpha_{t,i} \tag{19}$$

$$\hat{z}_t = \sum_i s_{t,i} a_i$$

where  $s_{t,i}$  is the binary variable introduced, which takes the value of 1 when region  $i$  is selected, and 0 otherwise. In order to use maximum likelihood estimation, we need to

marginalize the implicit variable and then take the lower bound of the logarithmic edge likelihood as the objective function  $L_s$ :

$$L_s = \sum_s p(s|a) \log p(y|s, a) \leq \log \sum_s p(s|a) p(y|s, a) = \log p(y|a) \quad (20)$$

where  $y$  is a description statement generated for image  $a$ . To find the gradient of the objective function:

$$\frac{\partial L_s}{\partial W} = \sum_s p(s|a) \left[ \frac{\partial \log p(y|s, a)}{\partial W} + \log p(y|s, a) \frac{\partial \log p(s|a)}{\partial W} \right] \quad (21)$$

Approximate with N Monte Carlo sampling method:

$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N} \sum_{n=1}^N \left[ \frac{\partial \log p(y|\tilde{s}^n, a)}{\partial W} + \log p(y|\tilde{s}^n, a) \frac{\partial \log p(\tilde{s}^n|a)}{\partial W} \right] \quad (22)$$

The moving average method is used to reduce the variance of the Monte Carlo method's estimated gradient. In the TH batch gradient descent, the moving average is estimated to be the cumulative sum of the previous logarithmic likelihood with exponential decay:

$$b_k = 0.9 \times b_{k-1} + 0.1 \times \log p(y|\tilde{s}_k, a) \quad (23)$$

To further reduce the variance, the entropy of multivariate Bernoulli distribution is introduced to improve the robustness of the algorithm:

$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N} \sum_{n=1}^N \left[ \frac{\partial \log p(y|\tilde{s}^n, a)}{\partial W} + \lambda_r (\log p(y|\tilde{s}^n, a) - b) \frac{\partial \log p(\tilde{s}^n|a)}{\partial W} + \lambda_e \frac{\partial H[\tilde{s}^n]}{\partial W} \right] \quad (24)$$

## 4. Experiments

### 4.1. Evaluation Criteria

In the test stage, BLEU [27] and CIDEr [28] were used in this experiment. BLEU measures the accuracy of the machine-generated description by using an n-gram to count the coverage between the generated description and the reference description. The greater the coverage, the more accurate the generated description. CIDEr performs TF-IDF weighting for each n-gram between sentences and calculates the cosine similarity between their TF-IDF weight vectors to measure the consistency between the reference description and generated description and evaluate the consistency and richness of the image description.

### 4.2. Experimental Details

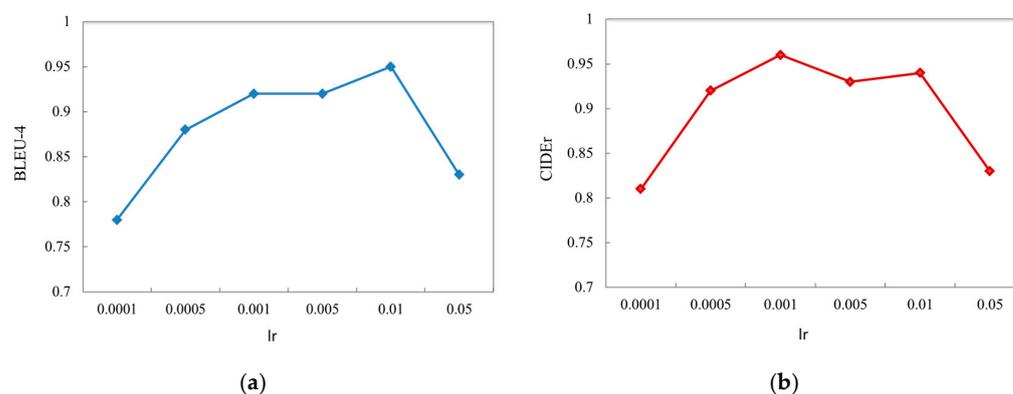
The platform used in the experiment was the GIGABYTE Technology RTX4090 24G with two graphics cards, and the deep learning framework was PyTorch. The loss function of the model adopted the cross-entropy loss function, and Adam was used as the algorithm optimizer to better train the model and improve the quality of the model. CUDA and cuDNN were used in training to accelerate the learning of GPU models and improve the training speed. Table 2 describes the hardware and software configurations.

**Table 2.** Hardware configuration table.

GPU	Video Storage	Frame	Programming Language
GIGABYTE Technology RTX4090	24G	pytorch	Python

In the dataset used in the experiment, there was a lot of noise in the pictures which would disturb and affect the experimental results. Therefore, it was necessary to preprocess the data before the experiment to reduce the impact of the original data on the experimental results and obtain real and reliable results. The specific preprocessing method was as follows: the input image was clipped, and the size of the training image in the dataset was adjusted to the input size and dimension of the ResNet50 network ( $224 \times 224 \times 3$ ); the diagnostic report statement corresponding to the image of the dataset was preprocessed; the characters, not letters or numbers, in the description statement were screened and removed; the words were divided by the space; and the words obtained after word segmentation were extracted and made into a vocabulary. The words that appeared fewer than five times were used as the <<UNK>> Mark and substitute, creating word-to-index and index-to-word mappings. To avoid the impact of excessively long description statements on the model, the maximum length of the description statements was set to 15 words.

In addition, before model training, it is necessary to determine the setting of the hyperparameter during training. The hyperparameter does not need to be driven by data, but needs to be adjusted manually before or during training. The learning rate of model training needs to be adjusted manually, and a too high or too low learning rate will lead to model optimization failure. In this paper, the model was trained by pre-setting different learning rates (Lr), and the evaluation results are shown in Figure 6. According to the experimental results, the learning rate with the highest score on the evaluation index was selected as the super parameter setting for the subsequent model training to better train the model.



**Figure 6.** (a) Impact of learning rate on BLEU-4 evaluation indicators; (b) impact of learning rate on CIDEr evaluation indicators.

#### 4.3. BLEU and CIDEr Score

In order to verify the effect of the proposed model on image description generation, three comparative experiments were undertaken in this part. Comparison experiment 1: The model presented in this paper was compared with baseline models such as Multimodal RNN, DeepVS, and Google NIC [28]. Each model was trained and tested on our spine X-ray dataset, and the test and comparison results are shown in Table 3.

Judging from the results, the model proposed in this paper was compared with Multimodal RNN, DeepVS, Google NIC, and Soft-Attention and Hard-Attention models using attention mechanisms that have a great improvement on the BLEU index to evaluate the smoothness and accuracy of generated sentences, and the CIDEr index to reflect the diversity of image description generation.

Comparison experiment 2: This model was compared with the models mentioned in the related work in Section 2 of this paper in our spinal X-ray image dataset. The comparison results are shown in Table 4.

**Table 3.** Comparison of the evaluation results of this model and the baseline model on the spinal X-ray image dataset.

Baseline Model	Datasets	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDeR
MultimodalRNN	Spine X-Ray	0.670	0.490	0.350	0.250	0.760
DeepVS	Spine X-Ray	0.625	0.450	0.321	0.230	0.660
Google NIC	Spine X-Ray	0.665	0.461	0.329	0.246	0.813
Soft-Attention	Spine X-Ray	0.707	0.492	0.344	0.243	0.802
Hard-Attention	Spine X-Ray	0.718	0.504	0.357	0.250	0.780
Ours	Spine X-Ray	0.763	0.592	0.445	0.360	0.847

**Table 4.** The evaluation results of this model are compared with those of other models in related work on spinal X-ray image datasets.

Methods of Related Work	Datasets	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDeR
Reference [17]	Spine X-Ray	0.695	0.390	0.384	0.291	0.812
Reference [20]	Spine X-Ray	0.789	0.483	0.365	0.259	0.451
Reference [21]	Spine X-Ray	0.673	0.489	0.355	0.259	0.766
Reference [23]	Spine X-Ray	0.467	0.361	0.298	0.257	0.808
Ours	Spine X-Ray	0.763	0.592	0.445	0.360	0.847

The comparison results on the spinal X-ray image dataset show that compared with the four models, the proposed model had considerable improvement in all evaluation indexes of BLEU-2, BLEU-3, BLEU-4, and CIDeR, except the BLEU-1 index.

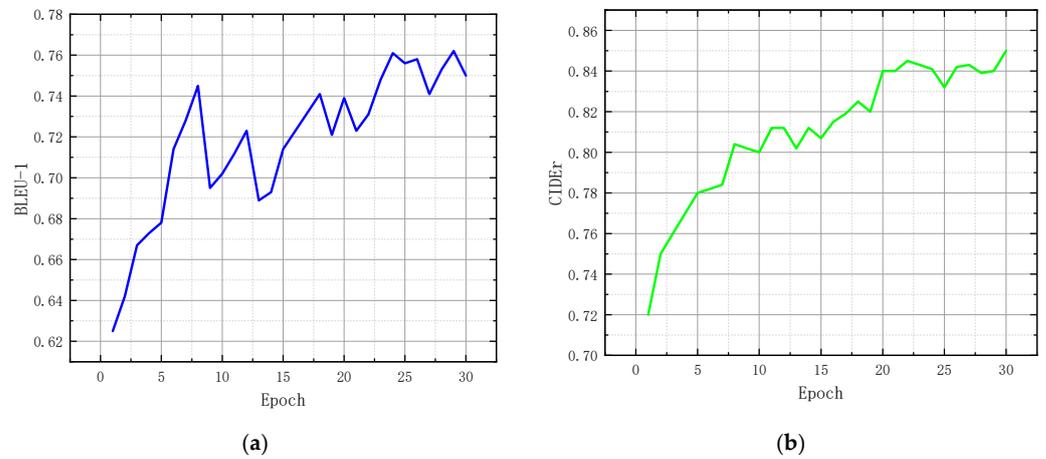
Comparison experiment 3: In order to demonstrate the effect of simplified explanatory sentences, we showed the BLEU and CIDeR scores of the trained model under the original diagnosis report and the simplified sentence report, respectively. The scores of BLEU and CIDeR are shown in Table 5. Non-scoliosis samples were not included in the test set because non-scoliosis images would automatically be described as empty sentences, perfectly repeating the original label, which would inflate the score.

**Table 5.** BLEU scores and CIDeR score for the spine in the test set, comparing models trained to reproduce the original report sentences vs. our simplified explanatory sentences.

N-Grams	Original Reports	Simplified Sentences
1-g	0.670	0.763
2-g	0.490	0.592
3-g	0.350	0.445
4-g	0.250	0.360
CIDeR	0.750	0.847

It can be seen from the table that the diagnostic text generated by the model trained with the simplified diagnostic report had much higher scores in both the BLEU1-4 index and the CIDeR index than that generated by the model trained with the original report, which confirms that simplifying complex diagnostic reports can improve the model's ability to reproduce sentences and obtain more accurate spinal image description. During

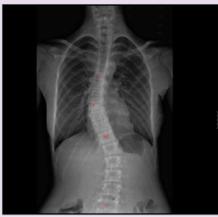
the training, the evaluation index changed with each iteration of the model, as shown in Figure 7. It can be seen from the figure that with each iteration of the model, the evaluation index constantly improved, and the model effect became better and better.



**Figure 7.** (a) BLEU-1 evaluation index change trend graph with iteration; (b) CIDEr evaluation index change trend graph with iteration.

#### 4.4. Description Results

To evaluate the descriptive ability of our model to generate sentences, 200 spinal X-ray images of different Lenke classifications were randomly selected from the test set for validation. The description of this model in the spinal X-ray image is shown in Figure 8.

			
<b>Original report:</b> No curvature and deformation was observed in the proximal thoracic and the cervical spine segment. The main thoracic segment was curved to the left. No obvious compression or collapse was observed in the cone. The thoracolumbar segment curved slightly to the right, the rib cone did not converge, the intercostal space was normal, and no other special signs were observed.	<b>Original report:</b> The spine was triple-curved, the cervical vertebra near the proximal thoracic segment was compressed, resulting in deformation, curved to the right, the main thoracic segment was curved to the left, the cone was intact, the cone space was normal, the lumbar regions segment was slightly curved to the right, the rib cone was not gathered, the intercostal space was normal, and other no special signs were observed.	<b>Original report:</b> The main thoracic is bent to the right with fixator shadows visible in the thoracic 1-7 vertebral body, while the lumbar regions is bent to the right with slight tapering of the lumbar 3-5 vertebral body. Multiple circular high-density shadows can be seen in the sternum area. The cervical spine slightly curved to the left, cervical spine presented recurve arch changes. observed.	<b>Original report:</b> With scoliosis deformity, no obvious curvature was observed in the thoracic vertebrae 1-8, 9-17 cone of the lumbar regions was clearly bent to the right, the cone of the right rib was gathered, the intercostal space was narrowed, the physiological curvature of the vertebral body in the cervical spine was straightened, and no other special signs were observed
<b>Generate report:</b> this spine has no curvature in the in the proximal thoracic, with curvature in the main thoracic, with curvature in the lumbar regions. it belongs to the Lenke3 type.	<b>Generate report:</b> The spine with curvature in the proximal thoracic and main thoracic. With curvature in the lumbar regions. It belongs to the Lenke4 type	<b>Generate report:</b> this spine has no curvature in the in the proximal thoracic, with curvature in the main thoracic, with curvature in the lumbar regions. it belongs to the Lenke3 type.	<b>Generate report:</b> There is no curvature in the main thoracic. With clear curvature in the lumbar regions. It belongs to the Lenke5 type.

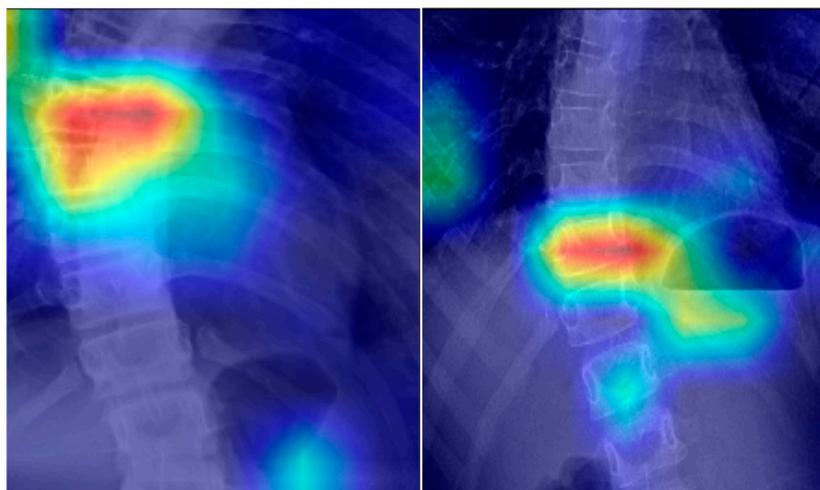
**Figure 8.** X-ray images of lenke1-lenke6 were randomly selected from the test set for testing. The original report of four samples is selected and the diagnostic text of this model is generated and displayed. The color-coded words are key features in the spine image. Different colors represent different locations.

It can be seen from the description results that the model in this paper has a good ability to describe the curvature characteristics of scoliosis. Different Lenke classification images can be identified, and the bending position, bending direction, and category of the image can be described in the description process. The sentences described had consistent sentence structures, as expected. In addition, we calculated the percentage of sentences containing spinal curvature positions and key feature descriptions in the models trained on the original diagnosis report and the models trained on the processed sentences, respectively, as shown in Table 6. It can be seen from the table that the model trained by the simplified sentence could generate a more accurate description of the disease characteristic information than the model trained by the original report.

**Table 6.** The percentage of explanatory sentences that contain an appropriate description of the location and character of the spinal curvature.

Original Reports	Simplified Sentences
73%	95%

Finally, in order to verify whether the spinal X-ray image description model could really pay attention to the semantics of key features in spinal images, attention mechanism visualization was adopted for the feature extraction process of the model, and it was generated into the corresponding thermal map output. The thermal map results are shown in Figure 9. It can be seen that the model was able to focus on the position of the curvature of the spine in the description, resulting in a more accurate and detailed description.



**Figure 9.** Two samples of our model visualizing attention in the description process.

#### 4.5. Doctor's Score

We selected 30 cases from the test set, which included a randomly selected Lenke classification image of scoliosis, randomly distributed from Lenke-1 to Lenke-6. For each Lenke classification, we used the model to generate the corresponding diagnostic text and gave it to five clinicians at Guizhou Orthopedics Hospital. We asked the doctors to compare our diagnostic text with the original diagnostic report and score the accuracy of the description of the key features of the spine and the accuracy of the Lenke classification of the spine. This was rated on a scale of 1 to 10, with 1 representing a completely unsatisfactory explanation and 10 representing a perfect explanation. Table 7 shows the average scores of the scoring results given by clinicians. As can be seen from the scoring results, the scores were mainly distributed between 8 and 9, so the diagnostic text generated by this model could be recognized by professional doctors.

**Table 7.** The scores given by a group of 5 doctors for different explanations for a diagnosis of spinal curvature, and the average score given by them.

Doctor	Score
Doctor1	8.8
Doctor2	8.3
Doctor3	8.5
Doctor4	9.1
Doctor5	8.6
Average	8.66

## 5. Discussion

To date, many researchers have made many attempts to generate descriptive texts from natural and medical images, but these methods have all used existing diagnostic report data as training labels, aiming to generate the same diagnostic text as the original report. However, the sentences in the original report were too complex, and the logic of the sentences was mostly affected by the subjective factors of clinicians. It is impossible for a computer to understand and reproduce the original report written by a human mind, so the diagnostic text generated by their model was not very effective.

In terms of spinal medical images, most of the current work is based on the disease detection and classification of spinal medical image datasets, and there has been no related research on the generation of spinal disease diagnosis reports. Therefore, this paper conducted a study on the automatic generation of a scoliosis diagnosis report. Different from the previous description methods, we created a new set of hand-marked sentences according to the visual features of spinal images under the guidance of professional radiologists as the labels for model training, and marked up the original diagnosis report into simple and consistent sentences with the same structure, using these sentences as a training model. Therefore, in the validation of the test sets, the model in this paper could output short but critical feature descriptions of the diagnostic text. In the model, the convolutional attention mechanism CBAM we added was different from the channel attention mechanism in traditional methods or the global attention mechanism. This attention mechanism could focus on channel and spatial features. Therefore, features with better correlation in spinal images could be extracted, and the description sentences generated could pay more attention to the information about spinal curvature. In addition, the attention mechanism was also added in the text decoding stage. The attention mechanism allowed the decoder to focus on the image region, with a high correlation with the words of the current moment, and it could predict the words of the current moment according to the words of the previous moment, instead of directly focusing on the whole image. Finally, the region features related to the current moment were used to generate words. This method was also the reason why the sentences generated by this model had good effects both in accuracy and in the score of evaluation indicators.

This method is not limited to models and datasets; it can also be applied to other medical image analysis tasks. However, there are limitations to this approach, as the preparation of short statement labels is cumbersome and a certain medical knowledge of scoliosis is required in the process of condensing traditional original diagnosis reports. Therefore, applying the simplified method of sentence labeling to the generation of other medical image diagnosis reports also requires a certain medical knowledge reserve or the guidance of professional doctors.

## 6. Conclusions

At present, before scoliosis surgery in hospitals, the diagnosis is made by the doctor observing the patient's spinal X-ray. The results of different doctors' observations may be biased, and the accuracy of the diagnosis report usually depends on the doctor's experience. Therefore, this paper proposes an end-to-end Lenke classification report automatic generation model for scoliosis that innovatively processes the original diagnosis

report into standardized simple sentences for training. In the model, image features are extracted by a deep residual network of fusion channels and spatial attention mechanisms. The improved M-LSTM long and short memory networks are integrated with the attention mechanism to decode the diagnostic text. Finally, in order to verify the effect of the model, the test was carried out on the scoliosis test set. From the perspective of the generated descriptive sentences, it can describe the curvature information of the spine in the spinal X-ray map through standardized and concise sentences and give the Lenke classification results of the spine. The generated descriptive sentences have achieved good scores on the BLEU and CIDEr indexes. In order to further prove the diagnostic significance of the generated diagnostic text, we gave the diagnostic text to five professional clinical doctors for scoring, and the average score was 8.66, which proved that the diagnostic text generated by our model can assist doctors in diagnosis.

**Author Contributions:** Y.T. designed the method, performed the experiments, and wrote the manuscript; Y.W. provided support in the data annotation; Z.H. advised in the design of the system and proofread the article; Q.W. methodology; X.W. Validation; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Guizhou Provincial Key Technology R&D Program [2022] General 264 and by the Guizhou Provincial Key Technology R&D Program [2021] General 442.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Private datasets were used in this study, and are not publicly available at this time.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Dash, S.K.; Acharya, S.; Pakray, P.; Das, R.; Gelbukh, A. Topic-based image caption generation. *Arab. J. Sci. Eng.* **2020**, *45*, 3025–3034. [[CrossRef](#)]
2. Allaouzi, I.; Ben Ahmed, M.; Benamrou, B.; Ouadouz, M. Automatic caption generation for medical images. In Proceedings of the 3rd International Conference on Smart City Applications, Tetouan, Morocco, 10–11 October 2018.
3. Shi, Z.; Zou, Z. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [[CrossRef](#)]
4. Ayesha, H.; Iqbal, S.; Tariq, M.; Abrar, M.; Sanaullah, M.; Abbas, I.; Rehman, A.; Niazi, M.F.K.; Hussain, S. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognit.* **2021**, *114*, 107856. [[CrossRef](#)]
5. Gale, W.; Oakden-Rayner, L.; Carneiro, G.; Bradley, A.P.; Palmer, L.J. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv* **2018**, arXiv:1806.00340.
6. Fu, K.; Jin, J.Q.; Cui, R.P.; Sha, F.; Zhang, C.S. Aligning where to see and what to tell: Image caption with region-based attention and scenespecific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2321–2334. [[CrossRef](#)] [[PubMed](#)]
7. 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. *arXiv* **2018**, arXiv:1801.04334.
8. Lenke, L.G. Adolescent idiopathic scoliosis: A new classification to determine extent of spinal arthrodesis. *J. Bone Jt. Surg. A* **2001**, *83*, 1169–1181. [[CrossRef](#)]
9. Wang, H.; Zhang, Y.; Yu, X. An overview of image caption generation methods. *Comput. Intell. Neurosci.* **2020**, *2020*, 214–217. [[CrossRef](#)] [[PubMed](#)]
10. Bai, S.; An, S. A survey on automatic image caption generation. *Neurocomputing* **2018**, *2018*, 291–304. [[CrossRef](#)]
11. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [[CrossRef](#)]
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
13. Melis, G.; Kočický, T.; Blunsom, P. Mogrifier lstm. *arXiv* **2019**, arXiv:1909.01792.
14. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *2015*, 212–218.
15. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image captioning: Transforming objects into words. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 232–236.

16. 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, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [[CrossRef](#)]
17. Shin, X.; Su, H.; Xing, F.; Liang, Y.; Qu, G. Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database for Automated Image Interpretation. *J. Mach. Learn. Res.* **2016**, *17*, 1–31.
18. Lin, M.; Chen, Q.; Yan, S. Network in Network, 2nd Int. Conf. Learn. Represent. ICLR 2014-Conf. *Track Proc.* In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings, Banff, AB, Canada, 14–16 April 2014; pp. 1–10.
19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [[CrossRef](#)]
20. Wu, L.; Wan, C.; Wu, Y.; Liu, J. Generative caption for diabetic retinopathy images. In Proceedings of the 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Shenzhen, China, 15–17 December 2018; pp. 515–519. [[CrossRef](#)]
21. Zeng, X.; Wen, L.; Liu, B.; Qi, X. Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing* **2019**, *392*, 132–141. [[CrossRef](#)]
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
23. Zhang, Z.; Xie, Y.; Xing, F.; McGough, M.; Yang, L. Mdnnet: A semantically and visually interpretable medical image diagnosis network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3549–3557. [[CrossRef](#)]
24. 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, Melbourne, Australia, 15–20 July 2018; pp. 2577–2586. [[CrossRef](#)]
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
27. 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 of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
28. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.