IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2021, accepted October 17, 2021, date of publication November 1, 2021,
date of current version November 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3124564

Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation

HYERYUN PARK[1,3], KYUNGMO KIM[1], SEONGKEUN PARK[2], AND JINWOOK CHOI[2,3,4]

[1]Interdisciplinary Program for Bioengineering, Graduate School, Seoul National University, Seoul 03080, South Korea
[2]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul 03080, South Korea
[3]Integrated Major in Innovative Medical Science, Graduate School, Seoul National University, Seoul 03080, South Korea
[4]Medical Research Center, Institute of Medical and Biological Engineering, Seoul National University, Seoul 03080, South Korea

Corresponding author: Jinwook Choi (jinchoi@snu.ac.kr)

This work was supported by the Grant from the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), by the Ministry of Health and Welfare, Republic of Korea, under Grant HI18C0316.

ABSTRACT The steadily increasing number of medical images places a tremendous burden on doctors, who toned to read and write reports. If an image captioning model could generate drafts of the reports from the corresponding images, the workload of doctors would be reduced, thereby saving time and expenses. The aim of this study was to develop a chest x-ray image captioning model that considers the differences between patient images and normal images, and uses hierarchical long short-term memory (LSTM) or a transformer as a decoder to generate reports. We investigated which feature representation method was the most appropriate for capturing the differences. The feature representations differed in terms of whether global average pooling was used for the visual feature vectors and how the feature difference vectors were generated. Experiments were conducted on two datasets using the proposed models and recent captioning models (X-LAN and X-Transformer). BLEU, METEOR, ROUGE-L, and CIDEr were used as evaluation metrics. The best model for most metric scores was the multi-difference non-average-pooling transformer model, which uses the transformer decoder, does not use global average pooling for the visual feature vectors, and applies the element-wise product. The transformer decoder was found to be more suitable than hierarchical LSTM. Furthermore, for models that do not condense features with global average pooling, the element-wise product was observed to be more effective than subtraction in expressing the feature differences.

INDEX TERMS Chest x-ray, deep learning, feature differences, medical image captioning.

I. INTRODUCTION

A. IMAGE CAPTIONING

Image captioning is a research field that focuses on methods for automatically generating text to explain the contents of an image. This area involves the convergence of computer vision to understand images and natural language processing to generate word sequences. Image captioning offers various applications, such as text-based image retrieval, related keyword assignment, human–robot interactions, and support for visually impaired people. Several methods have been developed for image captioning, including retrieval-based, template-based, and deep learning-based methods [1]–[7].

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

Most deep learning studies have used an encoder–decoder architecture with an attention mechanism [8]–[18]. The encoder transforms the image into a feature vector, whereas the decoder converts the feature vector into word sequences. We propose a new model that uses a convolutional neural network (CNN) as the encoder and hierarchical long short-term memory (LSTM) or a transformer [19] as the decoder. In recent years, deep learning-based methods have gained popularity in the image captioning field.

B. MEDICAL IMAGE CAPTIONING

Medical artificial intelligence applications are undergoing rapid expansion, from reading images to diagnosing diseases [20]. Image captioning has also been applied in the medical field (Fig. 1). As chest x-rays are the most common

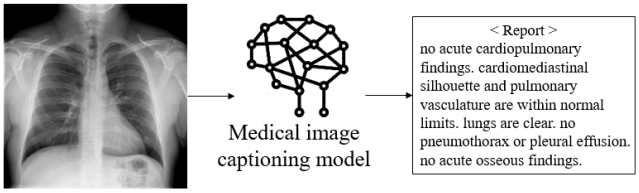

FIGURE 1. A medical image captioning model generates a draft report of the corresponding medical image.

types of medical images, and are important for screening and diagnosis, we conducted experiments using chest x-ray images. The number of medical images increases continually, which imposes a tremendous burden on doctors in terms of reading and writing reports. Medical image captioning can assist doctors by accelerating the reporting process and reducing their workload. However, little progress has been made in medical captioning compared to image captioning in other fields, and multiple factors have limited the performance of image captioning for chest x-rays.

The first challenge is that only a small number of publicly available datasets exist from previous chest x-ray image captioning studies: IU X-RAY, PEIR GROSS, ICLEF-CAPTION, and MIMIC-CXR [21]. This constrains research and development, and even PEIR GROSS and ICLEF-CAPTION provide captioning that is far from realistic because the images and captions are obtained from scientific articles. The recently released MIMIC-CXR [22] is the largest dataset that contains images, reports, and labels. Second, the datasets are skewed by the presence of numerous sentences that denote normal findings, but rare sentences with diverse contents regarding abnormal findings, leading to difficulties in learning. The third factor is clinical accuracy, particularly regarding the correct description of abnormal findings in the image. As chest x-ray images are very similar, it is important to identify and describe abnormal findings.

C. RELATED WORKS

Several studies have been conducted on chest x-ray image captioning. The model developed by Jing *et al.* [28] applies VGG-19 as a CNN to extract a visual feature vector from an image and uses a multi-layer perceptron for tag classification. The co-attention mechanism independently attends to the visual feature vector and tag embedding vector to create a context vector that is input into a hierarchical LSTM decoder. The TieNet [29] model consists of a ResNet-50 [30] and a flat LSTM decoder with a multi-level attention mechanism. This model can also classify disease labels based on the image features and text embeddings. Both the Jing and TieNet models were evaluated using the IU X-RAY dataset, and the Jing model exhibited superior performance. The HRGR-Agent [31] uses DenseNet or VGG-19 as a CNN and includes a sentence decoder that generates a topic state sequence. Given each topic state, a retrieval policy module decides to retrieve the template or to generate a sentence using the generation module. Both modules are updated by a reinforcement algorithm. The model that was developed by Srinivasan *et al.* [32] detects and crops lung regions using a single-shot multibox object detector, obtains visual features from a CNN, and creates tag embeddings. Two transformer-based decoders use the image and tag features to generate reports. The model that was developed by Liu *et al.* [33] uses DesNet-121 for the image encoder and hierarchical LSTM decoder, and applies self-critical sequence training [34] to consider both the readability and clinical accuracy. Lovelace and Mortazavi [35] established a model that extracts a visual feature vector by DesNet-121, and uses a transformer encoder and decoder. This model is trained using a language generation objective and clinical coherence objective to generate more clinically coherent reports. The model was evaluated on the MIMIC-CXR dataset. The model of Chen *et al.* [36] generates reports by means of a memory-driven transformer. It has a relational memory that records key information and incorporates this memory into the transformer decoder. The mDiTag(-) model [37] is similar to our approach; it investigates the differences between a patient image and a normal image, similar to when radiologists read images. This model utilizes multiple feature difference vectors, which are the result of subtracting the visual feature vectors of the normal image from those of the patient image. The model uses ResNet-152 as an encoder and hierarchical LSTM as a decoder.

New models have been progressively developed through recent studies on image captioning in the general domain. Faster R-CNN [23] is a model which is often used to extract salient region feature vectors. The GCN-LSTM model [24] leverages the interactions between objects by extracting regional representations from Faster R-CNN, and by constructing a spatial graph and a semantic graph to learn relation-aware region representations. The model uses these representations to generate text with two-layer LSTM as a decoder. The Reflective Decoding Network [25] uses the regional feature vectors of Faster R-CNN as the image input, and takes into account the word position information and distant words. Moreover, several studies have been conducted to enhance the attention mechanism. AoANet [26] adopts an AoA module that adds attention to consider the relevance between the original attention result and query. The AoA module is applied to the encoder and decoder of the image captioning model. The X-Linear Attention Network [27] uses the X-Linear attention block, which employs bilinear pooling for spatial and channel-wise bilinear attention. The X-Linear Attention Network uses regional feature vectors from Faster R-CNN as the input into the X-Linear attention blocks, and the output is passed to the LSTM and X-Linear attention block to generate word sequences.

D. AIM OF STUDY

As chest x-rays are part of the basic checkup routine in every hospital and clinic, the demand for reading x-rays has created difficulties. Moreover, because the number of chest x-ray images increases every year, the need to read and write

reports is a tremendous burden for doctors. If a deep learning model could generate drafts of reports to assist radiologists, their workload would be reduced, thereby saving time and expenses. The aim of this study was to develop a medical image captioning model using a CNN as an image encoder and by exploiting hierarchical LSTM or a transformer as the report generator. We investigated the simple question of which feature representation and method are the most appropriate for capturing the difference between patient and normal images. Using a given group of patient and normal images as the training input, we tested multiple feature difference vectors to generate draft reports.

Section II explains the baseline and the proposed models, whereas Section III describes the dataset and several experimental setups. The settings varied according to the selection of the decoder and the method used for feature representation. Section IV presents the outcomes of the metric evaluation, analysis of the model outputs, and the results for other datasets and models. Finally, the principal results and limitations are provided in Section V, and conclusions are drawn in Section VI.

II. METHODS

A. OVERVIEW OF MODELS

Three models consisting of an encoder–decoder architecture were compared. In each case, the encoder was ResNet-152, which is a CNN that extracts multiple visual feature vectors from an image. Two decoder types were used: a hierarchical LSTM with a co-attention mechanism and a transformer decoder. The decoder refers to multiple feature difference vectors that contain the differences between a patient image and a normal image, which are used to generate a report. Two selections were made regarding the feature representation method. The first was whether to use global average pooling on the visual feature vectors (Fig. 2). Through global average pooling, the three-dimensional visual feature vector of each layer becomes a one-dimensional vector by averaging the feature map. Without global average pooling, the three-dimensional visual feature vector is transformed into a two-dimensional vector by flattening the feature map. The fully connected layer adjusts the dimensions to 512 to match the decoder dimensions. The second selection was how to generate feature difference vectors: by subtracting the normal visual feature vectors from the patient visual feature vectors, or by multiplying these two vectors.

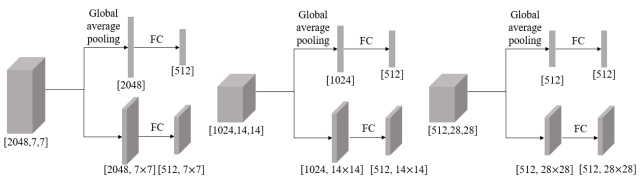

FIGURE 2. The difference according to whether global average pooling is used for visual feature vectors in each layer.

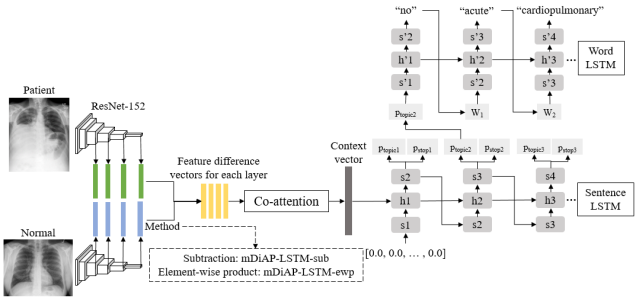

FIGURE 3. The mDiAP-LSTM model uses ResNet-152 to extract the visual feature vectors, applies global average pooling on these vectors, generates feature difference vectors by performing subtraction or the element-wise product, creates a context vector in every step of the sentence LSTM with a co-attention mechanism, and generates words using the word LSTM.

B. MODELS IN DETAIL

The baseline model was the multi-difference average pooling LSTM (mDiAP-LSTM) model, which was determined to be the best model in a previous study [37]. Fig. 3 presents the overall architecture of the mDiAP-LSTM model. The model uses ResNet-152 to extract the visual feature vectors from the final convolutional layer and three additional lower convolutional layers. Global average pooling was applied after extracting the four visual feature vectors from a patient image and a normal image, and feature difference vectors were generated. Consequently, the four feature difference vectors included information on the differences between the two images, which were input into the decoder. The hierarchical LSTM consisted of a sentence LSTM with a co-attention mechanism and a word LSTM that generated word sequences. The co-attention mechanism independently attended to the four feature difference vectors and produced a final context vector. The sentence LSTM used the final context vector to generate a topic vector and a stop vector. The word LSTM used the topic vector and embedding vector of the previous words as input, and predicted the next word at each step. The total loss was defined as the sum of the stop loss (binary cross-entropy loss) and word loss (cross-entropy loss).

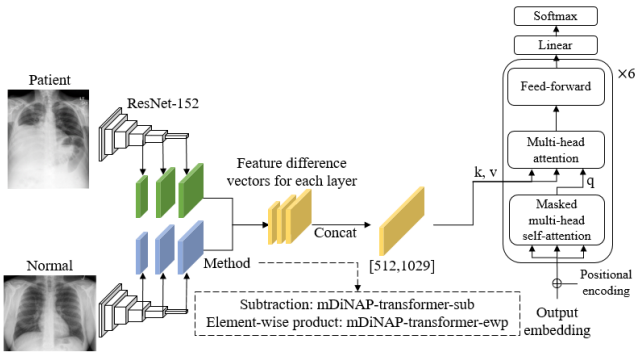

FIGURE 4. The mDiNAP-transformer model does not apply global average pooling for visual feature vectors, generates a concatenated feature difference vector, and sends this vector to the encoder–decoder multi-head attention of the transformer decoder as the key and value vectors.

The multi-difference non-average-pooling transformer (mDiNAP-transformer) model used a transformer decoder rather than the hierarchical LSTM (Fig. 4). The transformer decoder had six decoder blocks, each of which consisted of a masked multi-head self-attention, encoder–decoder multi-head attention, and a feed-forward layer. The mDiNAP-transformer model extracted three visual feature vectors: the final convolutional layer and two additional lower convolutional layers of ResNet-152. This model used three visual feature vectors instead of four owing to the number of parameters and complexity. The model did not apply global average pooling and generated three feature difference vectors. The concatenation of the three feature difference vectors resulted in 512 × 1,029-dimensional vectors. These vectors were used as the key and value vectors in the encoder–decoder multi-head attention part of every transformer block, and the query vectors were obtained from the layer below it. After passing through the six transformer decoder blocks, a linear layer and softmax layer computed the probabilities of all unique words to select a word. The loss function was the cross-entropy loss between the predicted and ground-truth words.

The multi-difference average-pooling transformer (mDiAP-transformer) model (Fig. 5) was similar to the mDiNAP-transformer model. Whereas the mDiNAP-transformer model used the visual feature vectors from three layers without global average pooling, the mDiAP-transformer model obtained the visual feature vectors from four layers with global average pooling. Subsequently, the mDiAP-transformer model created four feature difference vectors, the concatenation of which yielded 512 × 4-dimensional vectors. These vectors were used as the key and value vectors for the encoder–decoder multi-head attention of every transformer block. The other parts were the same as the mDiNAP-transformer model.

III. EXPERIMENTAL DETAILS

A. DATASET

We used the IU X-RAY dataset, which is accessible through the Open Access Biomedical Image Search Engine [38]. This dataset contains a series of chest x-ray images, reports, and tags. It includes 7,470 images, which are either posteroanterior or lateral views. Only 3,821 images of posteroanterior views were selected for this study. Every image has a corresponding report and tags. The report consists of comparison, indication, findings, and impression sections, and we only used the findings and impression sections as the model output. The Medical Text Indexer [39] program automatically extracts tags, which are Medical Subject Heading terms, from each report. A total of 210 unique tags exist, including ''normal,'' ''pleural effusion,'' ''lymph,'' ''hyperexpansion,'' ''mass lesion,'' and ''scarring.'' Among the images, 1,502 only had the ''normal'' tag; thus, we randomly sampled 75 images. The final dataset consisted of the same 2,394 image–report pairs as those from a previous study [37].

We also used a subset of the MIMIC-CXR dataset [40]. This dataset includes 1,083 images, corresponding reports, anatomical bounding boxes, audio, and eye gaze data. Images were sampled from MIMIC-CXR dataset for each diagnosis class, resulting in 360 images for pneumonia, 363 images for congestive heart failure, and 360 images for a normal status.

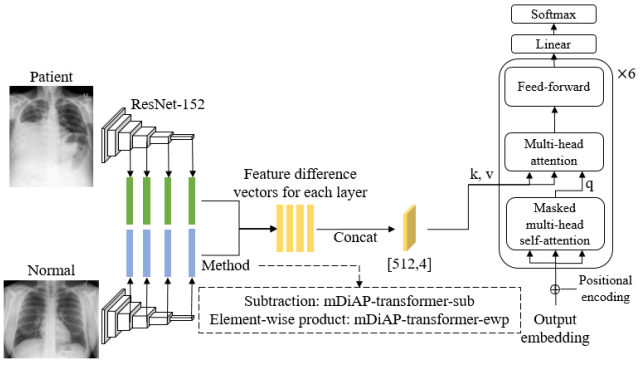

FIGURE 5. The mDiAP-transformer model applies global average pooling for the visual feature vectors, generates a concatenated feature difference vector, and sends this to the encoder–decoder multi-head attention of the transformer decoder as the key and value vectors.

B. EXPERIMENTAL SETUP

The decoder had an input vector dimension, a hidden vector dimension, and an output vector dimension of 512 for all models. The hierarchical LSTM decoder consisted of one layer of the sentence LSTM, which could generate a maximum of six sentences, and one layer of the word LSTM, which could generate a maximum of 36 words. The transformer decoder had six transformer layers and eight attention heads, and could generate up to 200 words. To prevent overfitting, dropout layers were included after each transformer sub-layer with a ratio of 0.1. The model used a regularization technique, namely label smoothing, and epsilon was set to 0.1. The transformer applied a beam search that maintained the most probable word sequences at each time step and finally selected the highest-probability sequence. The number of beams was four because most of the medical terms and the phrases indicating whether the findings were normal consisted of four or more words, such as ''no acute cardiopulmonary abnormality,'' ''degenerative changes are present,'' ''left lower lung opacity,'' and ''no pneumothorax or pleural effusion.'' The initial learning rate was 0.001 for the mDiAP-LSTM model, and 0.00004 for the mDiNAP-transformer and mDiAP-transformer models. All models used the Adam optimizer and ReduceLROnPlateau, which reduced the learning rate when the loss stopped decreasing. The total number of training epochs was 400 and the best model was defined as that with the maximum sum of all metric scores on the validation dataset. It took 1 day to run the mDiAP-LSTM model, 5 days to run the mDiNAP-transformer model, and 3 days to run the mDiAP-transformer model with a GeForce RTX 3090.

IV. RESULTS

A. METRIC EVALUATION OUTCOMES

The evaluation metrics were BLEU [41], METEOR [42], ROUGE-L [43], and CIDEr [44], which are word overlap measures. BLEU and METEOR evaluate the translation results, ROUGE-L measures the summarization performance, and CIDEr is a metric for image captioning. Table 1 displays the metric evaluation scores of the models for the IU X-RAY test dataset. The ''sub'' and ''ewp'' following the model name are abbreviations for subtraction and the element-wise product (i.e., means of creating a feature difference vector).

The mDiNAP-transformer-ewp model exhibitied the best performance on all metric scores. This model did not use global average pooling, used a transformer decoder, and applied the element-wise product to generate feature difference vectors. The models using the transformer decoder mainly outperformed the models with hierarchical LSTM. The results also demonstrated which feature representation method was better for the transformer decoder models. In the mDiNAP-transformer models, which did not use global average pooling and used non-condensed visual features, the element-wise product consistently prevailed in all metrics. However, in the mDiAP-transformer models, which used global average pooling and condensed visual features, the subtraction yielded better results for the BLEU and METEOR scores, whereas the element-wise product obtained better results in terms of the ROUGE-L and CIDEr scores.

B. MODEL OUTPUTS

Examples of the output from the best model, namely mDiNAP-transformer-ewp, are presented in Supplementary Material 1. There are six examples, each of which summarizes the model output of the test data. The ground truth of the first example is a normal report and the model output did not generate text regarding abnormal findings. In the second example, both the ground truth report and the model output indicate the presence of one abnormal finding (degenerative changes). The ground truth and model output of the third example both describe left base opacity, but the model does not mention a more specific cause (atelectasis/infiltration). The real report of the fourth example describes chronic obstructive pulmonary disease and aortic vascular calcifications, and the model output mentions emphysema associated with chronic obstructive pulmonary disease, but omits the calcifications. The actual report of the fifth example describes right pleural effusion, a blunted right costophrenic angle, and patchy left lower lobe airspace disease (infiltration). The model detected patchy left lower lobe airspace disease, but infiltration, which is the specific cause, is not mentioned. The model also states that bilateral pleural effusion exists, but only the right side exhibits pleural effusion. Moreover, the model incorrectly predicted atherosclerotic changes of the aorta and arthritic changes. In the sixth example, the ground truth report shows hyperexpanded lungs, which is suggestive of emphysema, right middle lobe airspace disease, which may be pneumonia, and degenerative changes. The model accurately identified hyperexpanded lungs, but omitted emphysema, and the model-generated report mentions focal airspace disease related to right middle lobe airspace disease, but omits detailed abnormal findings and pneumonia. Furthermore, the ground-truth report suggests the need for a follow-up examination after treatment, but the model does not.

C. EXTENDED EXPERIMENTS

We performed additional experiments to verify which feature representation method was the most appropriate for capturing the differences between normal and abnormal images. Two types of experiments were conducted: the first applied our models to the MIMIC-CXR dataset and the other adopted more powerful captioning models.

The experimental results for the MIMIC-CXR test dataset are presented in Table 2. Similar to the result of the IU X-RAY dataset, the mDiNAP-transformer-ewp model performed the best on all metric scores. The mDiNAP-transformer models outperformed the baseline mDiAP-LSTM models.

Further experiments were performed with more powerful captioning models, namely X-Linear Attention Networks (X-LAN) and X-Transformer [27]. Both models adopt the X-Linear attention block and use the feature difference vectors from each model as the input, as illustrated in Fig. 6 and Fig. 7. The X-LAN model consists of a transformer encoder with six transformer layers and an

TABLE 1. Metric evaluation results of models for IU X-RAY.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
mDiAP-LSTM-sub	0.3105	0.1868	0.1275	0.0894	0.1141	0.2456	0.1868
mDiAP-LSTM-ewp	0.2985	0.1875	0.1268	0.0883	0.1650	0.2733	0.2012
mDiNAP-transformer-sub	0.3355	0.1942	0.1219	0.0767	0.1519	0.2522	0.1775
mDiNAP-transformer-ewp	**0.3731**	**0.2260**	**0.1473**	**0.1010**	**0.1828**	**0.2930**	**0.3191**
mDiAP-transformer-sub	0.3616	0.2221	0.1439	0.0929	0.1621	0.2740	0.1935
mDiAP-transformer-ewp	0.3467	0.2157	0.1392	0.0901	0.1602	0.2949	0.2461

TABLE 2. Metric evaluation results of models for the MIMIC-CXR.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
mDiAP-LSTM-sub	0.2521	0.1671	0.1226	0.0921	0.1224	0.2669	0.1823
mDiAP-LSTM-ewp	0.2739	0.1740	0.1216	0.0897	0.1377	0.2684	0.1718
mDiNAP-transformer-sub	0.3177	0.1956	0.1309	0.0934	0.1480	0.2736	0.1913
mDiNAP-transformer-ewp	**0.3391**	**0.2166**	**0.1490**	**0.1071**	**0.1574**	**0.2900**	**0.2251**

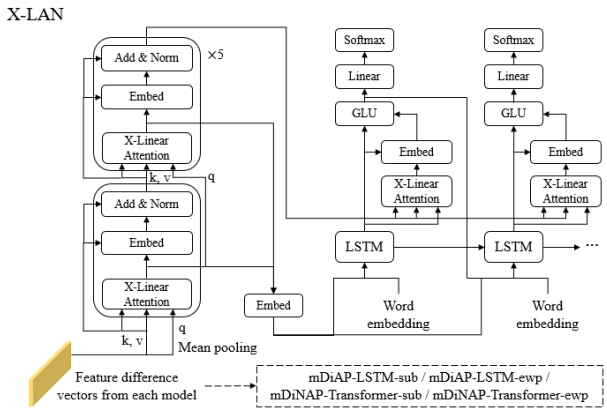

FIGURE 6. The X-LAN has a transformer encoder with X-Linear attention and a LSTM decoder with X-Linear attention.

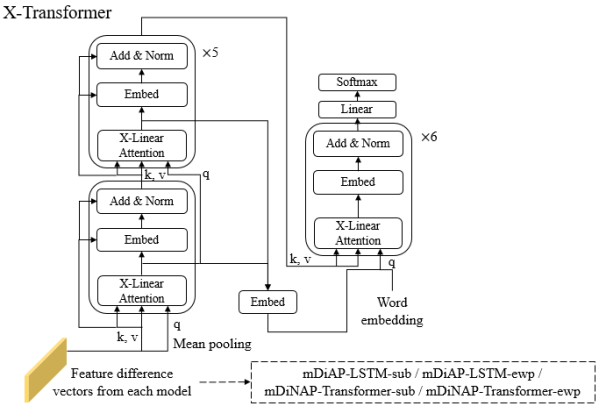

FIGURE 7. The X-Transformer has a transformer encoder with X-Linear attention and a transformer decoder with X-Linear attention.

LSTM decoder. The X-Transformer model has a transformer encoder and a transformer decoder, each with six transformer layers.

Table 3 displays the results of X-LAN and X-Transformer for the IU X-RAY dataset. In the case of X-LAN, when the feature difference vectors from the mDiNAP-transformer-ewp model were used, superior performance was observed in all metrics except for the ROUGE-L score. For X-Transformer, the performance was generally good when using the feature difference vectors from the mDiAP-LSTM-ewp model.

Table 4 presents the results for the MIMIC-CXR dataset. For both X-LAN and X-Transformer, better performance was generally achieved when using the feature difference vectors from the mDiNAP-transformer-ewp model.

In summary, leveraging the feature difference vectors of the mDiNAP-transformer-ewp model resulted in better performance in X-LAN and X-Transformer. In particular, taking advantage of non-condensed feature difference vectors was effective when applied to a smaller dataset, namely MIMIC-CXR. Furthermore, the element-wise product was

TABLE 3. X-LAN and X-Transformer results for IU X-RAY.

Model	Image Features	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
X-LAN	mDiAP-LSTM-sub	0.3619	0.2422	0.1633	0.1053	0.1870	**0.3356**	0.2875
	mDiAP-LSTM-ewp	0.3395	0.2324	0.1644	0.1130	0.1755	0.3324	0.2727
	mDiNAP-transformer-sub	0.3759	0.2384	0.1668	0.1220	0.1839	0.3171	0.3335
	mDiNAP-transformer-ewp	**0.3881**	**0.2451**	**0.1711**	**0.1249**	**0.1881**	0.3229	**0.3725**
X-Transformer	mDiAP-LSTM-sub	0.3413	0.2290	0.1642	0.1160	0.1734	**0.3370**	0.3618
	mDiAP-LSTM-ewp	**0.4286**	**0.2651**	**0.1787**	**0.1195**	**0.1986**	0.3226	0.3434
	mDiNAP-transformer-sub	0.3876	0.2222	0.1396	0.0952	0.1797	0.2956	0.3387
	mDiNAP-transformer-ewp	0.3627	0.2203	0.1491	0.1080	0.1803	0.3050	**0.4587**

TABLE 4. X-LAN and X-Transformer results for MIMIC-CXR.

Model	Image Features	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
X-LAN	mDiAP-LSTM-sub	0.2278	0.1694	0.1289	0.1002	0.1866	**0.3438**	**0.2181**
	mDiAP-LSTM-ewp	0.2669	0.1937	0.1459	0.1146	0.1978	0.3319	0.1946
	mDiNAP-transformer-sub	0.2778	0.1963	0.1469	0.1134	0.1842	0.3105	0.1231
	mDiNAP-transformer-ewp	**0.3102**	**0.2121**	**0.1558**	**0.1185**	**0.2098**	0.3254	0.2167
X-Transformer	mDiAP-LSTM-sub	0.3277	0.2214	0.1573	0.1174	0.2104	0.3278	0.2328
	mDiAP-LSTM-ewp	0.3277	0.2214	0.1573	0.1174	0.2104	0.3278	0.2328
	mDiNAP-transformer-sub	**0.3610**	0.2432	0.1745	0.1325	**0.2240**	0.3387	**0.4045**
	mDiNAP-transformer-ewp	0.3557	**0.2459**	**0.1778**	**0.1345**	0.2231	**0.3525**	0.4043

found to be more effective than subtraction when using non-condensed feature difference vectors. Condensed feature difference vectors are recommended when applying a strong captioning model such as X-Transformer with a larger dataset.

V. DISCUSSION

The metric results demonstrated that the transformer decoder was generally superior to the hierarchical LSTM as a decoder. We investigated which feature representation method was the most appropriate for capturing the differences between patient and normal images. If the model did not use global averaging pooling (i.e., did not condense the image features), the element-wise product predominated for most metrics. The element-wise product could maximize the difference by decreasing smaller values and increasing larger values. Not condensing means not averaging the features (i.e., using all specific features); thus, in this case, the element-wise product can yield even more benefits by means of detailed feature differences. This demonstrates that the element-wise product is effective in representing feature differences, when using more detailed image features. However, if the model condensed the image features using global average pooling, subtraction was superior in terms of the BLEU and METEOR scores, whereas the element-wise product was superior in terms of the ROUGE-L and CIDEr scores. As global average pooling involves averaging the feature map into a single value, the effect of the element-wise product may be weak. Therefore, in this case, it is difficult to determine whether either subtraction or the element-wise product is appropriate.

The model outputs revealed the strengths and weaknesses of the mDiNAP-transformer-ewp model. The model could detect opacity by observing the contrast, detect hyperexpanded lungs by determining the change in size, and identify the location of airspace disease. However, the model omitted the more specific cause of these abnormal findings. This demonstrates that by providing the visual feature vectors from the lower convolutional layer of ResNet-152, the model could detect changes in the contrast, size, and disease location, but failed to identify the more specific cause. Moreover, the poorly performing cases involved incorrect detection of the location or additional abnormal findings that were not present in the image being determined.

Examples of the outputs of all models are provided in Supplementary Material 2. Models with hierarchical LSTM generated numerous repetitive sentences, which is an ongoing problem in hierarchical LSTM. The mDiAP-LSTM-sub model repeated the sentence ''no pleural effusion or pneumothorax'' three times, and the mDiAP-LSTM-ewp model repeated ''there is no pneumothorax or pleural effusion'' three times. However, the boundaries of the sentences were very clear. Conversely, the models with a transformer decoder and beam search yielded few repetitive sentences, but the sentence boundaries were sometimes not clear. The mDiAP-Transformer-sub model outputs the sentence ''there is a calcified granuloma in the left upper lobe calcified granuloma,'' which is a fusion of ''there is a calcified granuloma in the left upper lobe'' and ''left upper lobe calcified granuloma.'' Furthermore, the last sentence did not end clearly. Similar phenomena were observed in the other models with the transformer.

Therefore, the principal limitation of this study is the difficulty of accurately describing all abnormal findings and detailed causes. In future work, the model can be improved by incorporating the classification results of the image to explain the more specific cause of the abnormal findings or applying other methods to represent the differences between the patient and normal images. Moreover, the proposal and use of a new medical term accuracy metric for training may improve the model. In certain cases, coherency is an issue. For example, a report may contain conflicting results, such as stating that pleural effusion is and is not present. The suggestion of a new coherency metric to penalize incoherent reports could be another direction for future research. The lack of evaluation

by a radiologist is also a limitation because the model results are not sufficient for analysis. Furthermore, research in the direction of measuring the reliability of each predicted sentence will make it easier for radiologists to revise the report.

VI. CONCLUSION

This study has proposed a chest x-ray image captioning model that generates draft reports of images. We experimentally investigated which feature representation method was the most appropriate for capturing the differences between patient and normal images. Overall, the best model was found to be the mDiNAP-transformer-ewp model, which used a transformer decoder to generate the report, did not use global average pooling for the visual feature vectors, and applied the element-wise product to generate feature difference vectors. The transformer decoder was more suitable than the hierarchical LSTM, and if the model did not condense features with global average pooling, the element-wise product was more effective than subtraction for expressing the feature differences. This model can assist doctors, thereby saving time and expenses, and it can also be extended to other medical images. In future research, the model should be improved in terms of clinical accuracy until it can be deployed for real-world medical applications.

REFERENCES

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 15–29, doi: 10.1007/978-3-642-15561-1_2.

[2] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, Aug. 2013, doi: 10.1613/jair.3994.

[3] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *Proc. 52nd Annu. Meeting Assoc. for Comput. Linguistics (Short Papers)*, vol. 2, 2014, pp. 592–598, doi: 10.3115/v1/P14-2097.

[4] A. Gupta, Y. Verma, and C. V. Jawahar, "Choosing linguistics over vision to describe images," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 606–612.

[5] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TreeTalk: Composition and compression of trees for image descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 351–362, Dec. 2014, doi: 10.1162/tacl_a_00188.

[6] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Lang. Learning.*, 2011, pp. 220–228.

[7] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 444–454.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2015, pp. 2048–2057.

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137, doi: 10.1109/CVPR.2015.7298932.

[11] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659, doi: 10.1109/CVPR.2016.503.

[12] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," in *Proc. Thematic Workshops ACM Multimedia Thematic Workshops*, 2017, pp. 305–313, doi: 10.1145/3126686.3126717.

[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086, doi: 10.1109/CVPR.2018.00636.

[14] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," *Pattern Recognit. Lett.*, vol. 143, pp. 43–49, Mar. 2021.

[15] H. Zhu, R. Wang, and X. Zhang, "Image captioning with dense fusion connection and improved stacked attention module," *Neural Process. Lett.*, vol. 53, no. 2, pp. 1101–1118, 2021.

[16] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.

[17] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Appl. Sci.*, vol. 8, no. 5, p. 739, May 2018.

[18] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao (2020 April), "Unified vision-language pre-training for image captioning and vqa," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, Apr. 2020, pp. 13041–13049.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.

[20] J. Choi, "AI in medicine: Need of orchestration for high-performance," *Healthcare Informat. Res.*, vol. 25, no. 3, p. 139, 2019.

[21] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in *Proc. 2nd Workshop Shortcomings Vis. Lang.*, 2019, pp. 26–36, doi: 10.18653/v1/W19-1803.

[22] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 91–99.

[24] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 684–699.

[25] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8888–8897.

[26] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4634–4643.

[27] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10971–10980.

[28] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2577–2586, doi: 10.18653/v1/P18-1240.

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[31] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, Dec. 2018, pp. 1537–1547.

[32] P. Srinivasan, D. Thapar, A. Bhavsar, and A. Nigam, "Hierarchical X-ray report generation via pathology tags and multi head attention," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2020.

[33] G. Liu, T. M. H. Hsu, M. McDermott, W. Boag, W. H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest X-ray report generation," in *Proc. Mach. Learn. Healthcare Conf.*, Oct. 2019, pp. 249–269.

[34] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.

[35] J. Lovelace and B. Mortazavi, "Learning to generate clinically coherent chest X-ray reports," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2020, pp. 1235–1243.
[36] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1439–1449.
[37] H. Park, K. Kim, J. Yoon, S. Park, and J. Choi, "Feature difference makes sense: A medical image captioning model exploiting feature difference and tag information," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2020, pp. 95–102.
[38] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2015.
[39] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers, "The NLM indexing initiative's medical text indexer," in *Proc. MEDINFO*, 2004, pp. 268–272.
[40] A. Karargyris, S. Kashyap, I. Lourentzou, J. T. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, and M. Moradi, "Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development," *Sci. Data*, vol. 8, no. 1, pp. 1–18, Dec. 2021, doi: 10.1038/s41597-021-00863-5.
[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.
[42] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Jun. 2005, pp. 65–72.
[43] L. Chin-Yew, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013
[44] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.

HYERYUN PARK received the B.S. degree from the Department of Software and the Department of Chemistry and Biological Engineering, Ajou University, South Korea, in 2019. Since 2019, she has been a Student in bioengineering and innovative medical science at Seoul National University. Her research interests include image captioning and natural language applications in medical field.

KYUNGMO KIM received the B.S. degree from the Department of Medical Information Technology Engineering, Soonchunhyang University, South Korea, in 2014, and the M.S. degree from the Department of Bioengineering, Seoul National University, where he is currently pursuing the Ph.D. degree. His research was on extraction of entities, including abbreviated words on medical documents. His interest is to develop text summarization models on various domains, including medical field.

SEONGKEUN PARK received the B.S. and M.D. degrees in medicine and the Ph.D. degree in biomedical engineering from Seoul National University, South Korea, in 1993 and 2010, respectively. He is currently a Researcher with the Department of Biomedical Engineering, Seoul National University Hospital. His research interests include biomedical signal processing and theories and techniques of artificial intelligence in medicine.

JINWOOK CHOI received the bachelor's degree in medicine and the master's and Ph.D. degrees in biomedical engineering from Seoul National University, Seoul, South Korea, in 1987, 1990, and 1993, respectively. Currently, he is an Associate Professor with the Department of Biomedical Engineering, College of Medicine, Seoul National University. His research interests include natural language applications in medical field, information extraction, automatic summarization, and application of HL7 CDA-based information sharing technologies.