Chest X-Ray Caption Generation with CheXNet

Vidura Wijerathna
Department of Computer Science
and Engineering
University of Moratuwa
Sri Lanka
vidura.prasangana.17@cse.mrt.ac.lk

Hemaka Raveen
Department of Computer Science
and Engineering
University of Moratuwa
Sri Lanka
raveenhansika.17@cse.mrt.ac.lk

Sachini Abeygunawardhana
Department of Computer Science
and Engineering
University of Moratuwa
Sri Lanka
deepasika.17@cse.mrt.ac.lk

Thanuja D. Ambegoda
Department of Computer Science
and Engineering
University of Moratuwa
Sri Lanka
thanujaa@uom.lk

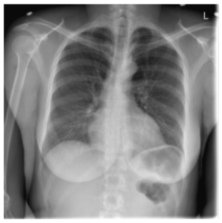
2022 Moratuwa Engineering Research Conference (MERCon) | 978-1-6654-8786-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/MERCon55799.2022.9906263

***Abstract*—Chest X-rays are provided with descriptive captions that summarize the crucial radiology findings in them in natural language. Although chest X-Ray image captioning is currently done manually by radiologists, automating it has received growing research interest in the medical domain because it is a tedious task and the high number of medical reports that are to be generated daily. In this paper, we propose an automatic chest X-ray captioning system consisting of two main components: an image feature extractor and a sentence generator. We did our experiment in two approaches. First, we tried using LXMERT, which is originally designed for question answering, as the sentence generator in our model combined with the Faster RCNN model. Second, we used CheXNet and a memory-driven transformer as the feature extractor and the sentence generator respectively. We trained and tested our model using the IU chest X-ray dataset. We evaluated the model using the BLUE, ROUGE-L and METEOR metrics which shows the CheXNet based approach outperforms the latter models.**

Keywords*—*Chest x-ray Captioning, Transformers, CheXNet

I. Introduction

The medical images extensively used to identify symptoms, signs of injury, and diseases are usually read by well-trained experts such as radiologists and physicians. However, with the increasing availability of medical images, now the radiologists and other physicians who are limited by speed and fatigue face difficulties in involving in the medical captioning themselves. The huge time taken for the task and the unreliable captions generated by inexperienced new physicians have become a bottleneck in the medical diagnostic and treatment pipeline. With that, the need for an effective and efficient method of captioning medical images has risen. An automated medical image captioning model will reduce the workload of humans, providing faster and more effective captions, which would increase the efficiency in the medical sector and speed up the process of diagnosis.



"The cardiomediastinal *silhouette is normal in size* and contour. No focal consolidation, pneumothorax or large pleural effusion. Negative for acute bone abnormality."

Fig. 1. An example of a human generated caption from the IU X-Ray dataset[2].

Although much prior work has focused on automating medical image captioning, it is not popular as a trustworthy solution yet. There are still multiple issues in the automation of medical image captioning to obtain results similar to human-generated captions. Chest X-ray captioning is considered a challenging task because it goes beyond identifying objects or any classification. The model should learn the connection between visual representations of the medical image and language semantics. This research project proposes an accurate machine learning-based Chest X-Ray captioning model.

This approach combines two main tasks: extracting the visual representations of a Chest X-ray image from a feature extractor and generating informative captions from a natural language processing technique. While many studies have been done on these tasks, transformers have recently become a trending methodology in language modeling. It is still novel to the chest x-ray image captioning domain as well. We did our experiments for the sentence generator in two approaches. First, we used the transformer-based LXMERT[17] model as our sentence generator, which is originally a question-

978-1-6654-8786-3/22/$31.00 ©2022 IEEE

answering model. We selected this model to try to take the advantage of the cross-modality encoder of LXMERT. However, we could not obtain good results. Then we used the R2Gen[1] model as our baseline model and modified it with the CheXNet[14] model. The best results were obtained by using the CheXNet model with the memory-driven transformer in the R2Gen[1] model. As our novel contribution, we used the CheXNet feature extraction model instead of the traditional CNN model, with the memory-driven transformer.

II. Related Work

A. Image Captioning

Initial studies[16] [11] [13] [23] of image captioning have used Convolutional Neural Networks (CNNs) and Recurrent Neural Network (RNN) as encoder and decoder respectively. Since CNNs are very wise in understanding images, they have been used for extracting features from images. Recurrent behavior of RNNs has given them the ability to remember previous inputs. Therefore RNNs are better at processing sequential data and they have been used for caption generation consuming features from images. Long short-term memory (LSTM) [5] networks are a special form of RNN with long-term memory and the ability to forget redundant memory. Since LSTM networks are better at understanding the semantics of the language, every RNN-based image captioning study has used it instead of a standard RNN decoder. After the introduction of transformers to the field, some recent work[18] has been focused on using transformers as the language modeling technology. Since Transformers have sequence-to-sequence architecture it is better at processing sequential data such as contextual text. In contrast to LSTM, Transformers consume inputs parallelly rather than one after one. Hence it could increase the efficiency of the data process and currently, transformers are mainly used in Natural Language Tasks.

B. Medical Image Captioning

Medical images are more complex than general images. Therefore various studies have done different things for models rather than general image captioning models. The study of Yin et al.,[22] and [10] has tried to replace a typical single RNN with a Hierarchical RNN (HRNN) which contains two LSTM-based RNNs for sentence topic creation and word generation. [13] has proposed an attention-based solution along with hierarchical RNN. And also the study used a “feature different vector” obtained from subtracting normal and patient images. [10] have proposed the REINFORCE algorithm to minimize the negative expected reward and a Clinically Coherent Reward (CCR) to optimize the generated clinical report for medical efficacy.

III. Method

In this section, we describe the architecture of the proposed model, each component, and its contribution to the model. The model consists of two main components, the image feature extractor, and the sentence generator.

Our experiments are done in two main approaches; using a completely new caption generation mechanism and improving an existing state-of-art model. In approach 1, we used LXMERT [17] which is originally an image-based question answering model. In approach 2 a memory-driven transformer[1] proposed by Chen et al, is used with CheXNet.

A. Approach 1 - LXMERT

Although plenty of transformer-based captioning models are currently used in many research domains, a minimal number of transformer-based studies have been done in the chest x-ray image captioning domain. Instead of using the typical transformer model to generate captions, we used the LXMERT model which is designed explicitly for question answering using images.

Since LXMERT is a question-answering model, we did several modifications to the model and tried to generate captions by it. The modified LXMERT architecture is given in Figure 2.

Similar to the original LXMERT, object bounding boxes identified from the images and their respective object features are input as the vision input in this approach as well. However, we provided captions as the language input, instead of the questions used in the original LXMERT. The first row of the language output which was just one key value and which is named cross-modality output by the authors was used to determine the predicted answer for the question in the original LXMERT. In contrast, we considered the whole 2D array for the caption generation as the caption is a sequence of words. The greedy search is used to extract words from the output vector. However, we did not receive any better results from this approach than the state-of-art. We realized that using captions as language input would not work well as the model tends to 100% depend on input caption and it would not learn anything from the object features.

B. Approach 2 - R2gen+CheXNet

1) CheXNet: The CheXNet [14] is a 121-layer Dense Convolutional Neural Network (DenseNet) trained on ChestX-ray 14 dataset [19] which contains 112,120 frontal-view chest X-ray images labelled with 14 diseases. Pneumonia detection is the main target behind the construction of CheXNet and it surpassed the radiologists’ performance on the F1 metric. Later authors extended the model to identify all 14 diseases from the dataset. The extended model has shown the state-of-art results in detecting diseases. We used the extended model in our implementation.

2) Caption Generation: The memory-driven transformer[1] developed by Chen at el. is used because it holds the state-of-arts results as a transformer-based caption generator. The model consists of two main modules, the encoder and the decoder, similar to the standard transformer. Relational Memory is a sub-module that is introduced to the decoder of the transformer to enable sharing of similar patterns in reports of similar images. A novel Memory-driven Conditional Layer Normalization (MCLN) has been proposed by Chen at el. to

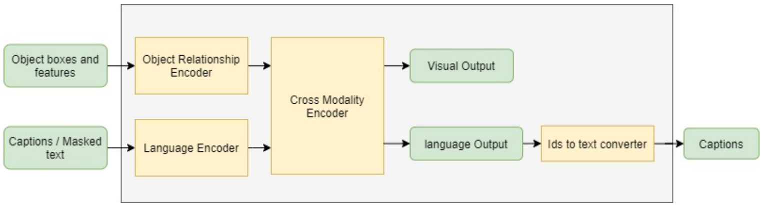


Fig. 2. LXMERT Captioning model architecture

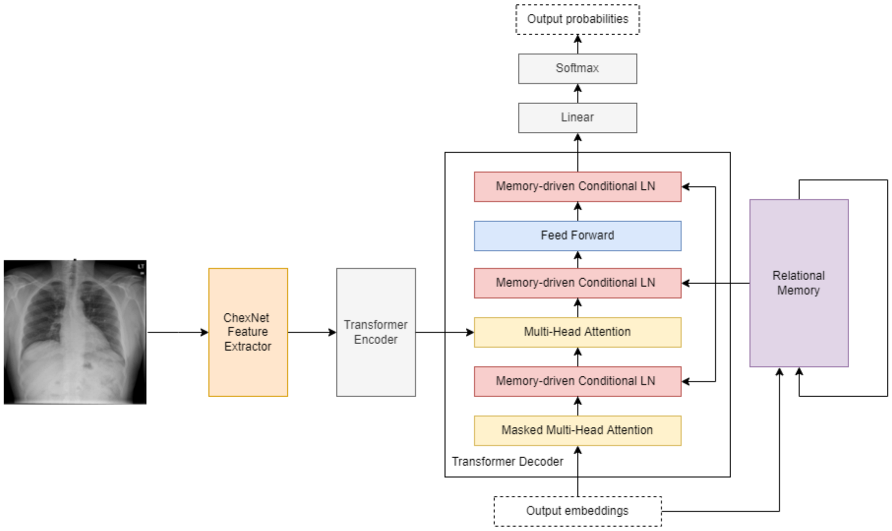


Fig. 3. CheXNet with Memory-driven Transformer architecture

bring the memory near to the decoder. MLCN has been used to incorporate Relational Memory and enhance the decoding of the Transformer.

In the second approach, the chest x-ray captioning model is implemented using the CheXNet feature extractor and the above mentioned memory-driven Transformer (Figure 3). We replaced the existing Resnet-101[4] model in R2Gen memory-driven transformer with the pre-trained CheXNet model to obtain better feature extraction ability. What motivated us for this modification is the lack of performance in chest x-ray image feature extraction of ResNet101[4] since it has been trained on ImageNet.

We used an already implemented CheXNet model in GitHub [21] which had better accuracy in labeling diseases than the paper [14]. The sequence-to-sequence paradigm is used in this method. X-ray images are used as a source to feature extractor. The CheXNet produces patch features of size 1024 while ResNet101[4] produces a size of 2048. Extracted patch features will go through the encoder and the decoder of the transformer respectively. The final output caption is constructed using beam search with a beam size of 3. Since it is possible to obtain better results by training this model on a classification chest x-ray dataset or a labeled chest x-ray dataset, we trained it using the Chest x-ray14 dataset. Chest x-ray14 is a labeled dataset with 14 diseases and the model named CheXNet is a DenseNet with 121 layers trained on the chest x-ray14 dataset.

After the modifications, the final model is trained on the IU X-RAY dataset for further improvement of the performance.

IV. Experiments

In this section, we discuss the datasets that have been used for experiments and implementation details of both approaches.

A. Dataset

IU X-RAY [2] dataset is used on all experiments done using the LXMERT model and Memory-driven Transformer. The dataset is collected by Indiana University. It contains 7,470 chest X-ray images along with 3,955 medical reports. Reports are available in four titles ”Findings”, ”Impression”, ”comparison” and ”indication”. We used captions provided under the ”Findings” title. Although it contains 3,955 reports, only 2,955 reports can be used as the rest of the reports have null values for ”Findings”. Table 1 shows the data splitting details of the IU X-RAY dataset for experiments with a Memory-Driven Transformer and table 2 shows the splits used for LXMERT.

B. Evaluation Metrics

The most common and reliable evaluation measures in biomedical image captioning, conventional natural language generation (NLG) metrics will be used to evaluate the performance of the models. The NLG metrics contain BLEU[12], ROUGE[9] and METEOR[3].

TABLE I. DATA SPLITTING OF IU X-RAY DATASET USED FOR MEMORY-DRIVEN TRANSFORMER

	Train	Validation	Test
Reports	2069	296	590
Images	4138	592	1180
Avg. Len.	32	31	28

TABLE II. DATA SPLITTING OF IU X-RAY DATASET USED FOR LXMERT

	Train	Test
Reports	2069	886
Images	4137	1773
Avg. Len.	31	31

BLEU - Bilingual Evaluation Understudy is worked by comparing system-generated text and reference text. The output score is always between 0 and 1. BLEU is a precision-oriented metric that relies on the ratio of word counts of system-generated text and common words in both reference and system-generated text.

ROUGE - Recall-Oriented Understudy for Gisting Evaluation is also a score based on the comparison of reference text and system-generated text. This is a recall-oriented metric that is calculated from the ratio between overlapping words and reference text word count.

METEOR - Metric for Evaluation of Translation with Explicit ORdering is also a score based on the unigram matching of reference text and system-generated text. combination of unigram-precision, unigram-recall, and a calculation of fragmentation that can identify how well-ordered the similar words in the system-generated text are in relation to the reference text

C. Implementation details

Two main experiments were done on LXMERT in the first approach. The visual input of the LXMERT model is object bounding boxes extracted from the images. Considering the recommendation of the LXMERT authors, we used Faster R-CNN[15] to acquire object bounding boxes from our chest x-ray images. Per each image, 36 bounding boxes with the highest confidence and their feature maps are extracted and used as the visual input to the LXMERT. The dimension of each feature set of objects was 2048. As the authors of LXMERT suggested, we used 9 language encoding layers, 5 cross-modality layers, and 5 object-relationship encoding layers for all experiments. Since all captions of images are domain-specific, tokenizing and word prediction has been done using a vocabulary file that contains 3572 words, extracted from findings IU X-RAY dataset using Bert Word Piece Tokenizer[20].

a) Experiment 1 using LXMERT: Object bounding boxes were used as vision input and the full caption was used as the language input in the training phase. The generated caption was acquired from the language output. The model started to over-fit from the first epoch unexpectedly and it tended to 100% depend on language input which is the caption. Fully masked text is used as captions for validating. Since the model is fully dependent on language input it always outputs the same caption when validating.

b) Experiment 2 using LXMERT: Everything was the same as in experiment 1 except the language input in the training phase. A text with a mask as every word ("[MASK] [MASK] [MASK] ... [MASK] [MASK]") was input in training as well as validation. Since we do not input any caption in training, the model does not use the relationship between caption and image objects using a Cross-Modality encoder. This experiment got better results than previous while predicted sentence and ground truth sentences had 29% common word percentage. Table III compares two test images with their ground truth captions and predictions got from LXMERT. The best results we acquired from the first approach are mentioned in Table V.

In the second approach, we replaced the traditional CNN model (ResNet101[4]) in the baseline[1] with a pre-trained CheXNet feature extractor. Output results of the proposed model were compared with the baseline[1] results. Even after the model reached the minimum value of validation loss, validation matrices showed improvements. Therefore, when choosing the best performing model we considered the highest BLEU-4 score instead of considering the minimum validation loss.

We trained the proposed model using the IU X-RAY dataset. Image caption (image -text) pairs are input to train the model. Furthermore, when we used the IU X-RAY dataset, we used two X-RAY images that belong to a particular patient with relevant captions as input. We used Adam optimizer [7] to train the proposed model. The optimizer is built using two input learning rate parameters: the learning rate for the visual extractor and the learning rate for the remaining parameters. The remaining parameters are the parameters of the proposed model that do not represent the parameters of the visual extractor. Based on these two learning rates, we experimented with the proposed model and obtained results in two ways.

c) Experiment 1 using CheXNet: The value of the learning rate for the visual extractor is set as 5e-5 and the value of the learning rate for the remaining parameters is set as 1e-4.

d) Experiment 2 using CheXNet: The value of the learning rate for the visual extractor is set as zero and the value of the learning rate for the remaining parameters is set as 1e-4. In this assignment, we allow only the memory-driven Transformer to learn from the IU X-RAY dataset in the training phase. CheXNet visual extractor is not allowed to learn.

According to the NLG Metrics, the best results were given in the 11th epoch. Predictions given by the proposed model are reported in TABLE IV. The NLG Metric for the proposed model is reported in TABLE V.

V. Analysis and Discussion

In the first approach, we tried to build a caption generation model using LXMERT which is originally a question-

TABLE III. COMPARISON OF GROUND TRUTHS AND PREDICTIONS GIVEN BY LXMERT. FINDINGS THAT ARE COMMON TO BOTH GROUND TRUTH AND PREDICTION ARE HIGHLIGHTED IN YELLOW.

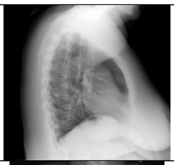
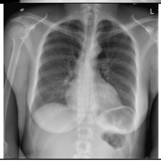
Image	Ground truth	Prediction
	"Cardiac and **mediastinal contours are within normal limits.** The lungs are clear. Acromioclavicular arthritis is present, XXXX severe."	"the and **mediastinal contours are within normal limits** . . . are . . bony structures are "
	"The cardiomediastinal **silhouette is normal in size** and contour. No focal consolidation, pneumothorax or large pleural effusion. Negative for acute bone abnormality."	"**the size silhouette normal** of vasculature not acute . . no size . . . . . no . . no . . "

TABLE IV. COMPARISON OF GROUND TRUTHS AND PREDICTIONS GIVEN BY BASE + CHEXNET. FINDING THAT ARE COMMON TO BOTH GROUND TRUTH AND PREDICTION ARE HIGHLIGHTED IN YELLOW

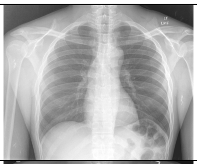
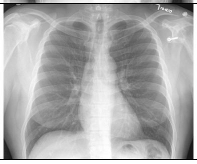
Image	Ground truth	Prediction
	"**the heart size** and mediastinal contours **appear within normal limits. no focal airspace consolidation pleural effusion or pneumothorax** . no acute bony abnormalities ."	"**heart size is normal** . **no focal airspace consolidations . no pneumothorax or pleural effusion** . no acute osseous findings ."
	"**the heart is normal in size. the mediastinum is unremarkable . the lungs are clear**."	"**the heart is normal in size. the mediastinum is unremarkable . the lungs are clear**."

TABLE V. THE PERFORMANCE COMPARISON OF PREVIOUS WORKS AND OUR MODELS ON IU X-RAY DATASET

Model	NLG Metrics					
	B-1	B-2	B-3	B-4	METEOR	ROUGE-L
HRGR-Agent [8]	0.438	0.298	0.208	0.151	-	0.322
CMAS-RL [6]	0.464	0.301	0.210	0.154	-	0.362
R2Gen [1]	0.47	0.304	0.219	0.165	0.187	0.371
LXMERT as caption decoder (Ours)	0.165	0.139	0.158	0.166	-	-
R2Gen + Chexnet (Ours: learning rate for the visual extractor as 5e-5)	0.448	0.289	0.211	0.164	0.185	0.356
R2Gen + Chexnet (Ours: learning rate for the visual extractor as zero)	**0.498**	**0.32**	**0.229**	**0.169**	**0.205**	**0.379**

answering model. It has shown good results for the question-answering task but the proposed model did not provide better results than our baseline for the caption generation task. According to our results, it is clear that the LXMERT highly depends on the language input. Therefore, the cross-modality encoder of LXMERT is not capable of understanding all relationships between captions and objects for captioning tasks. Therefore the proposed model could not provide better results than our baseline.

In the second approach, we could achieve better results compared to our baseline by freezing the weights of the CheXNet visual feature extractor. CheXNet is fine-tuned for extracting visual features of chest x-ray images. When training it with a captioning dataset (experiment 1 using CheXNet), its weights are changed in a manner to provide better caption predictions. However, when freezing its weights (experiment 2 using CheXNet), it could preserve image feature extraction ability and extract correct features throughout the whole training process. Therefore, fine-tuning the feature extractor doesn't improve the performance of the model. This transfers the sentence generation duty to the memory-driven transformer and helps to predict more accurate sentences according to image features. Also, we identified that DenseNet is a much better feature extractor for medical image captioning when compared to Resnet. Because our baseline has used Resnet-101 as the feature extractor to train the IU X-RAY dataset and it provided fewer results than our best model. Furthermore, we noted that pre-training the feature extractor on ChestX-ray images improves the performance of the model when compared to a pre-trained feature extractor which is trained on a general image dataset(ImageNet). Due to the above reasons, experiment 2 using the CheXNet approach could give better results than both baselines and experiment 1 using CheXNet.

Error Analysis: In the CheXNet pre-trained model, it is found that there is a class imbalance problem in the ChestX-ray 14 dataset [19]. This class imbalance is affected when training the CheXNet model on ChestX-ray 14 dataset. Therefore, in the future, we would like to address the data bias problem in the ChestX-ray 14 dataset and train the CheXNet model without class imbalance problems in the ChestX-ray 14 dataset. We expect that it will improve the accuracy of the chest x-ray image predictions.

VI. CONCLUSION

In this paper, we propose a Chest X-Ray Captioning model to automatically generate captions for chest x-rays. The experiment was done in two approaches on the IU chest x-ray dataset under BLUE, ROUGE-L, and METEOR metrics. The results of the first approach prove that using LXMERT in caption generation is not effective because it highly depends on its language input. The experimental results of the second approach demonstrate the effectiveness of the combination of ChexNet and a memory-driven transformer in chest x-ray caption generation. The scores achieved by our model show that it can be used to assist radiologists in clinical decision-making to reduce their workload. However, the accuracy of the generated captions by our model lags behind human-generated captions. More explorations are still needed in the chest x-ray captioning domain to obtain captions exactly similar to the human-generated captions.

REFERENCES

[1] Zhihong Chen et al. “Generating Radiology Reports via Memory-driven Transformer”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Nov. 2020, pp. 1439–1449. DOI: 10.18653/v1/2020.emnlp-main.112. URL: https://aclanthology.org/2020.emnlp-main.112.

[2] Dina Demner-Fushman et al. “Preparing a collection of radiology examinations for distribution and retrieval”. In: *Journal of the American Medical Informatics Association : JAMIA* 23 (July 2015). DOI: 10.1093/jamia/ocv080.

[3] Michael Denkowski and Alon Lavie. “Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems”. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 85–91. URL: https://aclanthology.org/W11-2107.

[4] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[5] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[6] Baoyu Jing, Zeya Wang, and Eric Xing. “Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 6570–6580. DOI: 10.18653/v1/P19-1657. URL: https://aclanthology.org/P19-1657.

[7] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2015).

[8] Yuan Li et al. “Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation”. In: *NeurIPS*. 2018, pp. 1537–1547. URL: http://papers.nips.cc/paper/7426-hybrid-retrieval-generation-reinforced-agent-for-medical-image-report-generation.

[9] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[10] Guanxiong Liu et al. “Clinically Accurate Chest X-Ray Report Generation”. In: *CoRR* abs/1904.02633 (2019). arXiv: 1904.02633. URL: http://arxiv.org/abs/1904.02633.

[11] David Lyndon, Ashnil Kumar, and Jinman Kim. “Neural Captioning for the ImageCLEF 2017 Medical Image Challenges”. In: *CLEF*. 2017.

[12] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[13] Hyeryun Park et al. “Feature Difference Makes Sense: A medical image captioning model exploiting feature difference and tag information”. In: Jan. 2020, pp. 95–102. DOI: 10.18653/v1/2020.acl-srw.14.

[14] Pranav Rajpurkar et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. In: *CoRR* abs/1711.05225 (2017). arXiv: 1711.05225. URL: http://arxiv.org/abs/1711.05225.

[15] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 91–99.

[16] Lakshminarasimhan Srinivasan and Dinesh Sreekanthan. “Image Captioning-A Deep Learning Approach”. In: 2018.

[17] Hao Tan and Mohit Bansal. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers”. In: *CoRR* abs/1908.07490 (2019). arXiv: 1908.07490. URL: http://arxiv.org/abs/1908.07490.

[18] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[19] Xiaosong Wang et al. “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *CoRR* abs/1705.02315 (2017). arXiv: 1705.02315. URL: http://arxiv.org/abs/1705.02315.

[20] *WordPiece tokenization - hugging face course*. URL: https://huggingface.co/course/chapter6/6.

[21] Nan Zhuang Xinyu Weng and Yingcheng Liu Jingjing Tian. *CheXNet for Classification and Localization of Thoracic Diseases*. https://github.com/arnoweng/CheXNet. 2017.

[22] Changchang Yin et al. “Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network”. In: *2019 IEEE International Conference on Data Mining (ICDM)*. 2019, pp. 728–737. DOI: 10.1109/ICDM.2019.00083.

[23] Dexin Zhao, Zhi Chang, and Shutao Guo. “A Multimodal Fusion Approach for Image Captioning”. In: *Neurocomputing* 329 (Nov. 2018). DOI: 10.1016/j.neucom.2018.11.004.