# MedSkip: Medical Report Generation Using Skip Connections and Integrated Attention

Esha Pahwa[1*] Dwij Mehta[2*] Sanjeet Kapadia[3*] Devansh Jain[4*] Achleshwar Luthra[5]

Birla Institute of Technology and Science, Pilani

{f20180675[1], f20190122[2], f20180137[3] ,f20180798[4] ,f20180401[5] } @ pilani.bits-pilani.ac.in

## Abstract

*Medical scans are extremely important for accurate diagnosis and treatment. To assist staff members in such crucial tasks, developing a computer vision model that efficiently processes a medical image and results in a generated report can be highly beneficial. Such a robust system can not only act as a helping hand for professionals but also eliminate the chances of error that might arise in the case of in-experienced staff members. However, previous studies lack focus on experimenting with the visual extractor, which is of eminent importance. Keeping this in mind, we propose a novel architecture of a modified HRNet which includes added skip connections along with convolutional block attention modules (CBAM). The entire architecture can be divided into two components, the first being the visual extractor where the pre-processed image is fed into the HRNet convolutional layers. Outputs of each down-sampled layer are concatenated after passing through the attention modules. The second component includes the use of a memory-driven Transformer that generates the report. We evaluate our model on two publicly available datasets, PEIR Gross and IU X-Ray, establishing new state-of-the-art for PEIR Gross while giving competitive results for IU X-Ray.*

## 1. Introduction

Medical images generated at pathology or radiology centers are used on a daily basis for accurate diagnosing of the disease or infection in a human's body. Almost every known disease requires laboratory evidence for confirmation and quick treatment. These scans are thus analysed by medical professionals and textual reports Figure 1 are created, which is often a tedious and time-consuming task. Given the number of patients in highly populated countries, the number of medical practitioners usually have to complete writing a pillar of reports in a limited amount of time. This can also lead to inaccurate diagnosis and thus can be harmful to the patient's life. Another issue arises when the medical practitioner has less experience, he or she may struggle to study the medical images, making the task extremely time-consuming.

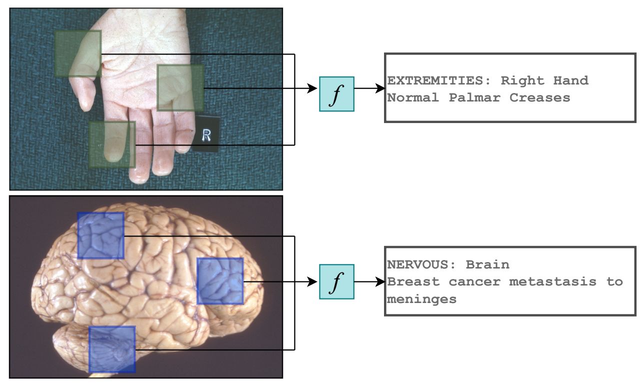


Figure 1: Sample images from the PEIR Gross dataset. Corresponding generated reports are shown alongside the image.

With the recent advancements in the field of artificial intelligence in developing state-of-the-art models for assistance in many day-to-day activities, deep learning would definitely be a promising approach to help pathologists and radiologists in diagnosing abnormalities and would also lessen their burden.

A complete medical report consists of a medical image along with a comprehensive explanation of the findings, impressions, abnormalities, and deductions. For example, as per [13] radiology reports should include narrative descriptions/itemization of findings, measurements, image annotations, key observations, inferences, and conclusions in addition to other components. These reports are complex and cannot be generated by the usual image captioning approaches since those are suitable for short sen-

[*]equal contribution

| Dataset Name | Year | #Images | Tags | #Reports | Average Sentence Length |
|---|---|---|---|---|---|
| IU X-Ray | 2015 | 7470 | MESH & MTI extracted terms | 3955 reports | 35 words |
| PEIR Gross | 2018 | 7442 | TF-IDF caption words | 7422 sentences | 12 words |

Table 1: Summary of Datasets used

tences. This problem is addressed by using a memory-driven Transformer that is capable of generating detailed reports. Another issue is localizing image regions that may contain abnormalities and specifically addressing these regions in the report [12]. This is tackled by using multiple attention modules linking the visual extractor to the Transformer.

Overall, the main contributions of our work are:

- We propose MEDSKIP, a novel visual extractor that incorporates skip connections and convolutional block attention modules with HRNet [22], combined with a memory-driven Transformer for medical report generation.
- We perform extensive experiments on two datasets to show the effectiveness of the proposed method.

The paper consists of the following subsections: Section 2 reviews related works. Section 3 introduces the basic methodology. Section 4 presents the experimental results and Section 5 concludes the paper.

## 2. Related Work

**Image Captioning** Image captioning is the task of automatically generating text descriptions for images. With the advent of deep learning, many works adopted a CNN-RNN framework [24], where CNNs were used to encode visual information and condition language generation while RNNs (LSTM [10]) were used as language models. The success of the attention concept [1] led to the addition of visual attention [8, 26] and visual as well as semantic attention [28] to the CNN-RNN architecture. Furthermore, Krause *et al*. [14] explored the use of hierarchical recurrent models to generate long paragraph captions.

More recently, the introduction of the Transformer [23] has led to the use of Transformer-based models for image captioning [9, 11, 4]. To balance features from the visual and textual modalities, Chen *et al*. [2] equipped an encoder-decoder attention mechanism with self-resurrecting activation units and leveraged pre-trained language models (BERT [7], GPT-2 [19]) for their linguistic knowledge.

**Medical Report Generation** Earlier works like [20] used a CNN-RNN framework that generated structured reports for chest X-ray images by predicting tags. Jing *et al*. [12] introduced a co-attention mechanism to localize abnormal regions, with a pre-trained VGG network [21] to get visual features and a hierarchical LSTM model to generate reports. Xue *et al*. [27] proposed a CNN-RNN architecture combined with an attention mechanism that used the encoding of an image and a generated sentence to guide the generation of the next sentence.

Liu *et al*. [16] introduced a domain-aware report generation system that first predicts the topics for the report and then conditionally generates sentences for these topics. The system is fine-tuned using reinforcement learning to improve the clinical accuracy of the generated reports.

Chen *et al*. [3] introduced a memory-driven Transformer with relational memory to record information from previous generation processes and a memory-driven conditional layer normalization to incorporate the relational memory into the Transformer.

Compared to previous studies, the approach proposed in this paper focuses on improving the extraction of visual features by adding skip connections and attention modules to the HRNet architecture, thereby leading to the generation of more accurate medical reports.

## 3. Proposed Methodology

This section highlights the main pathway followed by our model to learn pathology and radiology image datasets.

### 3.1. Visual Extractor

For a medical image $I$, the visual features $X$ are extracted using a visual extractor. We use a modified HRNet [22] with skip connections and convolutional block attention modules as our visual extractor, detailed below. The extracted features are then used as inputs by the Transformer.

**HRNet** Ke Sun *et al*. [22] introduced HRNet for the human pose estimation task. HRNet starts with a high-resolution branch in the first stage. In every following stage, a new branch is added to current branches in parallel with $\frac{1}{2}$ of the lowest resolution in current branches. As the network has more stages, it will have more parallel branches with different resolutions, and resolutions from previous stages are all preserved in later stages. It has performed extremely well on semantic segmentation, instance segmentation, and

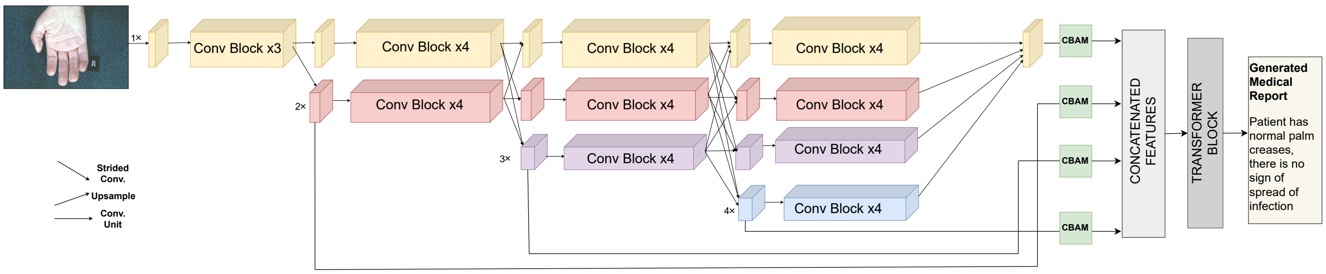


Figure 2: Block diagram of the modified HRNet architecture. The output of each downsampling layer is extracted and passed through CBAM modules (as depicted in green boxes). Results from each attention block are concatenated and fed into the Transformer.

object detection tasks. We have used a modified pose HRNet as our visual extractor. We have made two changes to HRNet's standard architecture.

- **Skip Connections:** The residual branches of HRNet with lower resolution run parallel to the branch with the highest resolution. As each downsampling layer is encountered, we extract the feature representation of the downsampled block.

- **Attention Module:** The extracted features are then fed into a simple convolutional block attention module which tries to extract the most important features from a feature representation. Once these features have been passed through the attention modules, they are concatenated together to create a 2048 dimensional feature vector.

**Convolutional Block Attention Modules** To inculcate attention into our work, we used the convolutional block attention module proposed by Sanghyun Woo *et al*. [25]. We use this because the module can be used as an additional plugin to our skip connections and it is end to end trainable. The module can be divided into two parts which are spatial and channel attention submodules. Given an intermediate feature map $\boldsymbol{F} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ as input, CBAM sequentially infers a 1D channel attention map $\boldsymbol{M_c} \in \mathbb{R}^{\mathrm{C}\times 1\times 1}$ and a 2D spatial attention map $\boldsymbol{M_S} \in \mathbb{R}^{1\times\mathrm{H}\times\mathrm{W}}$. The overall transformation performed by the module can be summarized as:

$$\boldsymbol{F'} = \boldsymbol{M_c}(\boldsymbol{F}) \otimes \boldsymbol{F} \tag{1}$$

$$\boldsymbol{F''} = \boldsymbol{M_s}(\boldsymbol{F'}) \otimes \boldsymbol{F'} \tag{2}$$

where $\otimes$ denotes element-wise multiplication. $\boldsymbol{F''}$ is the final refined output after being processed by the attention module.

### 3.2. Transformer

We adopt the Transformer model introduced by Vaswani *et al*. [23]. Transformer is an encoder-decoder model where the encoder contains stacked layers of self-attention and feed-forward neural network, and the decoder uses self-attention on words and cross-attention over the output of the last encoder layer.

**Encoder** We use the standard encoder from Transformer that operates directly on the visual features extracted by the Visual Extractor 3.1. The encoding process can be formalized as:

$$\{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_S}\} = f_e(x_1, x_2, ..., x_S) \tag{3}$$

where the outputs are the hidden states $h_i$ encoded from the visual features $x_i$ from the visual extractor and $f_e(.)$ refers to the encoder.

**Decoder** We use a modified version of Transformer's decoder introduced by Chen *et al*. [3]. The modified decoder contains a relational memory (RM) to facilitate learning from patterns in reports and record key information of the generation process. Further, a memory-driven conditional layer normalization (MCLN) is proposed to incorporate relational memory into the decoder. We refer the reader to Chen *et al*. [3] for a detailed description of the memory-driven decoder.

## 4. Experimental Results

The complete architecture of the CNN-Transformer network used is given in Figure 2.

### 4.1. Dataset Details

- PEIR GROSS: The dataset was first introduced in [12] wherein images were downloaded from the official

| Dataset | Model | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE |
|---|---|---|---|---|---|---|---|
| PEIR Gross | LRCN [8] | 0.261 | 0.184 | 0.136 | 0.088 | 0.135 | 0.254 |
| | SOFT ATT [26] | 0.283 | 0.212 | 0.163 | 0.113 | 0.147 | 0.271 |
| | MLC+COATTENTION+LSTM [12] | 0.300 | 0.218 | 0.165 | 0.113 | 0.149 | 0.279 |
| | R2GEN [3] | 0.379 | 0.262 | 0.193 | 0.130 | 0.191 | 0.411 |
| | HRNET | 0.312 | 0.189 | 0.132 | 0.083 | 0.126 | 0.308 |
| | MEDSKIP(Ours) | 0.399 | 0.278 | 0.209 | 0.148 | 0.176 | 0.414 |
| | MEDSKIP + CBAM(Ours) | 0.389 | 0.268 | 0.201 | 0.141 | 0.166 | 0.395 |
| IU X-Ray | LRCN [8] | 0.369 | 0.229 | 0.149 | 0.099 | 0.155 | 0.278 |
| | SOFT ATT [26] | 0.399 | 0.251 | 0.168 | 0.118 | 0.167 | 0.323 |
| | R2GEN [3](Avg. of 3 runs) | 0.428 | 0.272 | 0.196 | 0.151 | 0.177 | 0.350 |
| | HRNET | 0.427 | 0.251 | 0.177 | 0.134 | 0.173 | 0.338 |
| | MEDSKIP(Ours) | 0.467 | 0.297 | 0.214 | 0.162 | 0.187 | 0.355 |
| | MEDSKIP + CBAM(Ours) | 0.467 | 0.303 | 0.210 | 0.155 | 0.197 | 0.371 |

Table 2: Comparison of the full model results on PEIR GROSS dataset (upper part) and IU X-Ray (lower part) with previous works. Red denotes the best results and Blue represents the next highest value. The last 3 visual extractor models from both datasets are the models that we have trained ourselves. R2GEN results were replicated using their code. The remaining represent replicated results reported by Jing *et al*. [12].

website, also a digital library called The Pathology Education Informational Resource (PEIR) [1]. Each image is of 528 x 792 resolution. It consists of publicly accessible 7442 teaching images, spread across 21 predefined subcategories. As mentioned in [12], the vocabulary size of the total image captions is 4,452. It is different from IU-XRAY as it has single sentences as captions, unlike a report. Each image on average contains a 12 word caption.

- IU X-Ray: The dataset [5] can be publicly accessed through the Open Access Biomedical Image Search Engine [2] . It consists of 7,470 frontal and lateral chest X-rays along with their radiology report are divided into four sections. The 'comparison' section lists previous information about the patient such as preceding medical exams; the 'indication' section contains symptoms or reasons of examination; the 'findings' section contains detailed radiology observations; and the 'impression' section outlines the final diagnosis. [18] reports that 104 reports contained no image, 489 were missing 'findings', 6 were missing 'impression', and 25 were missing both 'findings' and 'impression'. Overall, it contains a total of 6674 training instances, the rest being used for testing purposes.

The train-test-validation split for the PEIR Gross dataset is 72:8:20 and for the IU X-Ray dataset is 70:10:20. The details about the datasets have been summarized in Table 1 and in the supplementary material.

[1]https://peir.path.uab.edu/wiki/Main_page
[2]https://openi.nlm.nih.gov

### 4.2. Experimental settings

Images are resized to $256 \times 256$ dimensions and then random cropping is performed to bring down the size to $224 \times 224$. After performing randomized horizontal flipping, the images are normalized, grouped into batches, and are fed into the model. The model is compiled using PyTorch and a single Tesla K80 GPU has been used for training. The model has been trained for 20 epochs with a batch size of 16 for each variation. The hyperparameters for training have been chosen after conducting extensive experiments. The training is done using cross-entropy loss with the ADAM optimizer and the learning rate as 1e-4 for all the parameters. The value of $\beta_1$ and $\beta_2$, i.e. the parameters used for calculating the moving average of the gradients and its square are 0.999 and 0.9 respectively. The number of layers of the Transformer is 3 and beam search has been used as the sampling method. The remaining hyperparameters of the Transformer are the same as used in [3].

### 4.3. Evaluation Metrics

The performance of the aforementioned models is evaluated using BLEU (Papineni *et al.* [17]), METEOR (Denkowski and Lavie [6]) and ROUGE-L (Lin [15]) metrics.

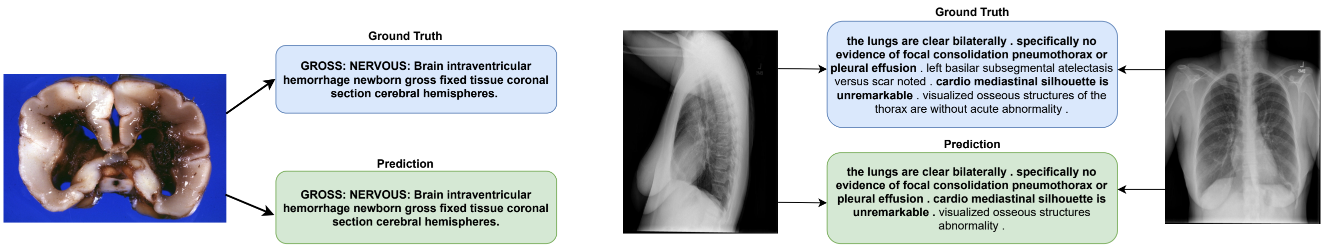


Figure 3: Visual results for MEDSKIP for PEIR Gross and IU X-Ray datasets respectively. Green box displays the prediction whereas Blue box depicts the ground truth report.

### 4.4. Performance Evaluation

Table 2 summarises the results obtained by the different visual extractor backbones on PEIR Gross and IU X-Ray respectively. Our baseline network corresponds to HRNet, while our proposed network called MEDSKIP includes the modifications of added skip connections. Additional experiments with skip connections and integrated attention modules (CBAM) were performed. Additional experiments conducted on the HRNet baseline highlight the positive influence of skip connections and attention modules. Training the model on MEDSKIP resulted in a score of 0.399 BLEU-1 and 0.278 BELU-2 on PEIR Gross test dataset. This alone beat the current state-of-the-art algorithms on the PEIR Gross dataset. Similarly for the IU X-Ray dataset, the model was able to achieve a BLEU-1 test score of 0.467 with MEDSKIP. Furthermore, all the variations using HRNet were tested similarly.

MEDSKIP outperforms HRNet for both datasets. However, CBAM doesn't generalise well over different datasets. Specifically, for PEIR Gross, MedSkip + CBAM performs worse due to the different types of images (and different body parts) present in the dataset. On the other hand, MedSkip + CBAM outperforms MedSkip for IU X-Ray. The reason for this could be the fact that each report in IU X-Ray contains two images, thus CBAM has more information to attend on. Further, all the images in IU X-Ray are uniform (chest x-rays) and make it easier for the model to generalise.

### 4.5. Discussion and Comparison

We compare our models with those in previous studies, including conventional image captioning models as well as models proposed specifically for medical report generation. The results are reported in Table 2 for PEIR Gross and IU X-Ray. These results show that improved visual features can lead to better-generated reports.

Works such as LRCN [8] and Soft Att [26] are specifically used for generating short sentences, however using a simple HRNet alone surpasses their results owing to its dense layers. In [3], visual features have been provided little importance, whereas in our work additional network attributes have been introduced, results of which are reflected in Table 2. For PEIR Gross, as compared to [12] CoAtt module formed by combining both visual and semantic features, we extract visual features alone from each downsampled branch which helps in accomodating significant areas in the images of the diverse range of affected body parts. To take it one step further, in IU-XRAY dataset, CBAM is needed to identify differences between similarly shaped grayscaled chest images. It can be seen that slightly better results for BLEU-1, BLEU-2, METEOR and ROUGE metrics are obtained for the same with the integrated attention modules. As opposed to the hierarchical LSTM used in [12], the memory driven transformer used [3] also helps in scoring an efficient and effective approach towards generating longer reports (as compared to the image captioning task).

## 5. Conclusion

In this study, we propose MEDSKIP network consisting of a modified HRNet and added skip connections. Convolutional Block Attention Modules were also integrated as part of the visual extractor which helps the model learn specific features of the medical image. This is followed by a memory-driven Transformer which gives us our generated report. A significant increase is observed in the case of PEIR Gross dataset, which contains pathology images that are not limited to just one organ where the model beats the previous state-of-the-art values for all six metrics. For IU X-Ray, it is able to achieve competitive results compared to the previous state-of-the-art values. It can be seen that MedSkip is capable of generalizing for all medical images including body parts and radiology images.

## 6. Acknowledgments

We want to thank the members of Computer Vision Research Society, BITS Pilani (CVRS[3]) for their helpful suggestions and feedback.

[3]https://sites.google.com/view/thecvrs

## References


[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

[2] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient image captioning by balancing visual input and linguistic knowledge from pretraining. *arXiv preprint arXiv:2102.10407*, 2021.

[3] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Nov. 2020.

[4] Marcella Cornia, Matteo Stefanini, L. Baraldi, and R. Cucchiara. Meshed-memory transformer for image captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10575–10584, 2020.

[5] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[6] 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*, pages 85–91, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

[7] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[8] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014.

[9] Simao Herdade, Armin Kappeler, K. Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, 2019.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[11] Lun Huang, Wenmin Wang, J. Chen, and Xiao-Yong Wei. Attention on attention for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642, 2019.

[12] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[13] Charles E. Kahn, Curtis P. Langlotz, Elizabeth S. Burnside, John A. Carrino, David S. Channin, David M. Hovsepian, and Daniel L. Rubin. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856, 2009. PMID: 19717755.

[14] J. Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3345, 2017.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[16] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. *CoRR*, abs/1904.02633, 2019.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[18] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, 2019.

[19] Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[20] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and R. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506, 2016.

[21] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[24] Oriol Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[25] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[26] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine*

*Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[27] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 457–466, Cham, 2018. Springer International Publishing.

[28] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, 2016.