**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 are created, which is 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.

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-
<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Year</th>
<th>#Images</th>
<th>Tags</th>
<th>#Reports</th>
<th>Average Sentence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>IU X-Ray</td>
<td>2015</td>
<td>7470</td>
<td>MESH &amp; MTI extracted terms</td>
<td>3955 reports</td>
<td>35 words</td>
</tr>
<tr>
<td>PEIR Gross</td>
<td>2018</td>
<td>7442</td>
<td>TF-IDF caption words</td>
<td>7422 sentences</td>
<td>12 words</td>
</tr>
</tbody>
</table>

Table 1: Summary of Datasets used

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
Convoluted 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 $F \in \mathbb{R}^{C \times H \times W}$ as input, CBAM sequentially infers a 1D channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$. The overall transformation performed by the module can be summarized as:

$$F' = M_c(F) \otimes F$$ (1)

$$F'' = M_s(F') \otimes F'$$ (2)

where $\otimes$ denotes element-wise multiplication. $F''$ 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:

$$\{h_1, h_2, ..., h_S\} = f_e(x_1, x_2, ..., x_S)$$ (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
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEIR Gross</td>
<td>LRCN [8]</td>
<td>0.261</td>
<td>0.184</td>
<td>0.136</td>
<td>0.088</td>
<td>0.135</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>SOFT ATT [26]</td>
<td>0.283</td>
<td>0.212</td>
<td>0.163</td>
<td>0.113</td>
<td>0.147</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td>MLC+COATTENTION+LSTM [12]</td>
<td>0.300</td>
<td>0.218</td>
<td>0.165</td>
<td>0.113</td>
<td>0.149</td>
<td>0.279</td>
</tr>
<tr>
<td></td>
<td>R2GEN [3]</td>
<td>0.379</td>
<td>0.262</td>
<td>0.193</td>
<td>0.130</td>
<td>0.191</td>
<td>0.411</td>
</tr>
<tr>
<td></td>
<td>HRNET</td>
<td>0.312</td>
<td>0.189</td>
<td>0.132</td>
<td>0.083</td>
<td>0.126</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td>MEDSkip(Ours)</td>
<td>0.399</td>
<td>0.278</td>
<td>0.209</td>
<td>0.148</td>
<td>0.176</td>
<td>0.414</td>
</tr>
<tr>
<td></td>
<td>MEDSkip + CBAM(Ours)</td>
<td>0.389</td>
<td>0.268</td>
<td>0.201</td>
<td>0.141</td>
<td>0.166</td>
<td>0.395</td>
</tr>
<tr>
<td>IU X-Ray</td>
<td>LRCN [8]</td>
<td>0.369</td>
<td>0.229</td>
<td>0.149</td>
<td>0.099</td>
<td>0.155</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>SOFT ATT [26]</td>
<td>0.399</td>
<td>0.251</td>
<td>0.168</td>
<td>0.118</td>
<td>0.167</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>R2GEN [3](Avg. of 3 runs)</td>
<td>0.428</td>
<td>0.272</td>
<td>0.196</td>
<td>0.151</td>
<td>0.177</td>
<td>0.350</td>
</tr>
<tr>
<td></td>
<td>HRNET</td>
<td>0.427</td>
<td>0.251</td>
<td>0.177</td>
<td>0.134</td>
<td>0.173</td>
<td>0.338</td>
</tr>
<tr>
<td></td>
<td>MEDSkip(Ours)</td>
<td>0.467</td>
<td>0.297</td>
<td>0.214</td>
<td>0.162</td>
<td>0.187</td>
<td>0.355</td>
</tr>
<tr>
<td></td>
<td>MEDSkip + CBAM(Ours)</td>
<td>0.467</td>
<td>0.303</td>
<td>0.210</td>
<td>0.155</td>
<td>0.197</td>
<td>0.371</td>
</tr>
</tbody>
</table>

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].

4.2. Experimental settings

Images are resized to 256 × 256 dimensions and then random cropping is performed to bring down the size to 224 × 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.
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 down-sampled branch which helps in accommodating 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 grayscale 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 in [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 (CVRS3) for their helpful suggestions and feedback.

---

3https://sites.google.com/view/thecvrs
References


