Literature Review for Deep Learning for Automatic Generation of Medical Image Report

Hendawi Bayan, Master student at the Data Science program at the HSE university

***Abstract*— Artificial intelligence has spread rapidly in various areas of daily life, including applications found in mobile phones and self-driving cars, as well as a remarkable spread in the medical field, such as early detection of cancerous tumors in the body and the use of robotic arms in surgical operations, as well as the need to understand and read medical radiographic images and automatically generated medical reports.**

***Keywords*— Deep Learning, Transformer, LSTM, RNN, CNN.**

I. INTRODUCTION

The radiologist reads the X-ray images by looking at the medical image and determining what the pathological condition is, and this task takes between 10 and 20 minutes in furthermore to the effort required due to the large number of x rays that they should always read on a daily, that also increased as epidemics spread, and we have seen in the recent pandemic (COVID-19), and other mistakes in the medical report due to insufficient experience.

One of the applications of image captioning, which is one of the most significant and difficult tasks in deep learning, is the generating of medical reports. Image captioning is the process of writing a description of an image; it is a simple operation for us as humans, but we should transfer this technique to the computer for it to grasp the image's information and write it automatically. This requires the use of two models:

- The first model extracts important features from the image by interpreting its content from its colors, edges, and forms.
- The second model generates text from the useful features extracted from the prior model.

II. RELATED WORK

We can divide the related work into two parts image captioning and medical image captioning.

A. Image Captioning

Xu et al. [1] use CNN as the encoder, converting the image to a fixed-length vector representation known as the image feature map, which is then sent to the Decoder, which is a long short-term memory that generates one word at a time based on a context vector, the previous hidden state, and previously generated words.

You et al. [2] deploy a Convolutional Neural Network (CNN) to extract a top-down visual feature and detect visual concepts at the same time. Then, an Recurrent Neural Network (RNN) that generates the image caption, uses a semantic attention model to link the visual feature with visual concepts.

Fu et al. [3] present a new modeling contribution that considers scene-specific contexts, visual features are extracted from the image after it has been analyzed and represented with various visual regions. The visual feature vectors are then input into a Long-Short Term Memory network (LSTM), which uses the transition of visual attention to predict both the sequence of concentrating on various locations and the sequence of generating words. A scene vector, a global visual context taken from the entire image, also governs the neural network model.

Yao et al. [4] initially identify a group of important image regions by the faster R-CNN. Then, on the discovered regions, a semantic/spatial network with directed edges is created, with the vertex representing each region and the edge denoting the semantic/spatial link between them. The structured semantic/spatial graph is then used to contextually encode regions with visual relationships using Graph Convolutional Networks (GCN). Following that, each kind of graph's learned relation aware region-level features is input into a single individual attention LSTM decoder for sentence production. They employed two decoders during the inference stage.

Fang et al. [5] suggest a multi-layer LSTM-based deep attention language model that can learn more abstract word information, as well as three overlapping approaches for generating attention context vectors.

Yang et al. [6] consider CNN and RNN encoders with RNN decoders. The review network conducts a series of review steps on the encoder hidden states using an attention mechanism, and after each review step, it outputs a thought vector; the thought vectors are utilized as the input of the attention mechanism in the decoder.

B. Medical Image Captioning (Generating Report)

Wu et al. [7] combine CNN to produce captions, RNN is used. The LSTM model and word embeddings are paired with a CNN image embedder. An image is encoded to create a vector, which the CNN will send to the LSTM. The LSTM then decodes the vector to produce a sequence.

Lyndon et al. [8] do as follows: captions that previously had numerous sentences become a single sentence after pre-processing (deleting punctuation marks from captions, converting captions to lower case, removing stop-words, and applying stemming). Pre-processed images are sent through an InceptionV3 CNN, which creates image embedding, which is then passed to the LSTM as an initial state only, and not utilized in further time steps. Following an initial state, each state's LSTM creates output, which is then passed on to a word-embedding layer and subsequently to a Softmax layer, which calculates the probability of the generated word in the dictionary. The model is trained in two stages. The weights of CNN are fixed in the first phase, but only the LSTM is trained. The complete model is then trained end-to-end.

Liang et al. [9] use CNN and LSTM to produce natural language phrases after training three deep learning models. To choose which deep model to utilize for predicting image caption, a three-class SVM classifier is trained at the same time. In addition, as a complement to the final caption, the Nearest Neighbor method is used to get a comparable image and its caption from the training data.

Su et al. [10] use an encoder-decoder system to create captions. They examine two convolutional neural network (CNN) architectures for the encoder: ResNet-152 and VGG-19 and employ the LSTM recurrent neural network for the decoder. The mechanism of attention is also addressed.

Spinks et al. [11] use the adversarial regularized autoencoder ARAE model to generate a textual representation of all captions. A CNN is used to map each picture to the continuous representation space in a second phase.

Hasan et al. [12] employ the VGGnet-19 deep CNN model with fine-tuning on the given ImageCLEF training dataset to extract the image feature representation. To create a caption, the decoder employs the LSTM network with a soft attention mechanism, which predicted one word at each time step based on a context vector.

Liu et al. [13] introduce the hierarchical generation strategy with a CNN-RNN-RNN architecture, CNN extracts visual features, which are then pooled to provide an average pooled vector, which is then put into Sentence-LSTM to create topics and an end token to stop the Word-LSTM. Word-LSTM is used to produce the word using the topic vector and pooled vector. The generated word is used to calculate attention. Finally, all the created words are concatenated to form a sentence. Duplicate sentences are found in the produced report and are eliminated during post-processing.

Xue et al. [14] propose a multimodal recurrent generation model with attention on radiology reports. The pre-trained resnet-152 image encoder extracted global visual features automatically. A single-layer LSTM predicted the first word of the sentence after a global visual feature vector is input into a sentence generating model that works as a sentence decoder. Using this sentence LSTM, the sentence is created word by word. The generative model for recurrent paragraphs consists of a sentence encoder and an attentional sentence decoder generated paragraphs sentence by sentence. A Bi-directional LSTM and a 1D encoder are employed as sentence encoders.

Xu et al. [15] use the CNN-RNN model framework, which included an attention mechanism. The encoder, which is based on the pre-trained ResNet-101, extracts visual features from the input images. LSTM network in the decoder to generate a caption by creating one word at each time step based on a context vector that captures the visual information.

Baoyu Jingy et al. [16] present two models: co-attention mechanism and hierarchical LSTM, the first of which is used to localize sub-regions in medical images to generate text that is compatible with regions, and the second of which is used to generate long reports.

Huang J et al. [17] extract features from medical images using a pre-trained CNN model and create reports using an LSTM model.

Li et al. [18] employ ResNet-121 as an encoder to create a feature map that could be applied to either the cropped or original image. To create sentences, feature maps are input into a visual attention-based LSTM.

Christy Y. Li et al. [19] propose a new approach called Knowledge-driven Encoder, Retrieve, and Paraphrase (KERP), in which they use CNN to extract features from medical images, then pass them to an encode module, which converts them into a structured abnormality graph, then to a retrieve module, which retrieves text based on the abnormalities detected previously, and finally to a paragraph module, which rewrites the paragraph in a way that is appropriated for the specific case.

Changchang Yin et al. [20] use a CNN with global pooling for abnormality detection and an HRNN (Hierarchical Recurrent Neural Network) for caption prediction. The HRNN has two levels: sentence RNN with attention to generating topic vector by LSTM and word RNN to generate meaningful paragraph based on the topic vector by LSTM. And also, they suggest a subject matching system that lowers the number of sentences duplicated in a single report.

The strong transformer is used by Omar Alfarghaly et al. [21]. They extract features and tags from medical images using a pre-trained Chexnet model, then use predicted tags to compute weighted semantic features, and lastly construct a medical report by tuning the visual and semantic features of a pre-trained GPT2 model.

Liu et al. [22] propose the Medical-VLBERT model. The visual context of the input images is first encoded as spatial features using a convolutional neural network (CNN). The predefined terminology embeddings are then associated with visual contexts and medical textbook embeddings, resulting in visual-terminological and textual-terminological features, respectively. The terminology encoder is defined in the BERT visual language (VLBERT).

III. Discussion and conclusion

All prior techniques are not relevant to real-world applications, because of different issues in existing models that need to be addressed and there is still a lot of potentials to improve the performance of creating radiological reports. The following are some difficulties of current approaches and prospective future research initiatives to bridge the gap between academic research and real-world radiology report generation:

- The quantity of publicly accessible datasets for medical image captioning is limited and unbalanced, there are no reports against many images.
- The prior approaches are uninterpretable, therefore the radiologist cannot rely on them in the real world.

Future work will focus on creating a large-scale dataset in collaboration with domain experts, as well as developing an interpretable and industry-standard automatic report generation system that will generate a complete medical report tailored to radiologists' needs.

References

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell Neural image caption generation with visual attention. In ICML, pages 2048–2057, 2015.

[2] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.503.

[3] Fu, K., Jin, J., Cui, R., Sha, F., & Zhang, C. (2017). Aligning were to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2321–2334. https://doi.org/10.1109/tpami.2016.2642953.

[4] Yao, T., Pan, Y., Li, Y., & Mei, T. (2018). Exploring visual relationships for image captioning. *Computer Vision – ECCV 2018*, 711–727. https://doi.org/10.1007/978-3-030-01264-9_42.

[5] Fang, F., Wang, H., Chen, Y., & Tang, P. (2018). Looking deeper and transferring attention for image captioning. *Multimedia Tools and Applications*, *77*(23), 31159–31175. https://doi.org/10.1007/s11042-018-6228-6.

[6] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. Salakhutdinov. Review networks for caption generation. In NIPS, pages 2361–2369, 2016.

[7] Wu, L., Wan, C., Wu, Y., & Liu, J. (2017). Generative Caption for Diabetic Retinopathy Images. *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*. https://doi.org/10.1109/spac.2017.8304332.

[8] D. Lyndon, A. Kumar, J. Kim, Neural captioning for the ImageCLEF 2017 medical image challenges, CEUR Workshop Proc. 1866 (2017).

[9] S. Liang, X. Li, Y. Zhu, X. Li, S. Jiang, ISIA at the ImageCLEF 2017 image caption task, CEUR Workshop Proc. 1866 (2017).

[10] Y. Su, F. Liu, M.P. Rosen, UMass at ImageCLEF caption prediction 2018 task, CEUR Workshop Proc. 2125 (2018).

[11] G. Spinks, M.F. Moens, Generating text from images in a smooth representation space, CEUR Workshop Proc. 2125 (2018).

[12] S.A. Hasan, Y. Ling, J. Liu, R. Sreenivasan, S. Anand, T.R. Arora, V. Datla, K. Lee, A. Qadir, C. Swisher, O. Farri, PRNA at ImageCLEF 2017 caption prediction and concept detection tasks, CEUR Workshop Proc. 1866 (2017).

[13] G. Liu, T.-M.H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically Accurate Chest X-Ray Report Generation, (2019). http://arxiv.org/abs/1904.02633.

[14] Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G. R., & Huang, X. (2018). A multimodal recurrent model with attention for Automated Radiology Report generation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018*, 457–466. https://doi.org/10.1007/978-3-030-00928-1_52.

[15] J. Xu, W. Liu, C. Liu, Y. Wang, Y. Chi, X. Xie, X. Hua, Concept detection based on multilabel classification and image captioning approach - DAMO at ImageCLEF 2019, CEUR Workshop Proc. 2380 (2019) 9–12.

[16] Baoyu, J., Pengtao , X., & Eric , P. X. (n.d.). On the Automatic Generation of Medical Imaging Reports. ArXiv. https://doi.org/arXiv:1711.08195.

[17] Huang, J.-H., Huck Yang, C.-H., Liu, F., Tian, M., Liu, Y.-C., Wu, T.-W., Lin, I.-H., Wang, K., Morikawa, H., Chang, H., Tegner, J., & Worring, M. (2021). DeepOpht: Medical report generation for retinal images via deep models and visual explanation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/wacv48630.2021.00249.

[18] X. Li, R. Cao, D. Zhu, Vispi: Automatic Visual Perception and Interpretation of Chest X-rays, (2019). http://arxiv.org/abs/1906.05190.

[19] Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for Medical Image Report Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 6666–6673. https://doi.org/10.1609/aaai.v33i01.33016666.

[20] Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., & Zheng, Q. (2019). Automatic Generation of Medical Imaging Diagnostic Report with a hierarchical recurrent neural network. 2019 IEEE International Conference on Data Mining (ICDM). https://doi.org/10.1109/icdm.2019.00083.

[21] Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., & Fahmy, A. (2021). Automated radiology report generation using conditioned Transformers. Informatics in Medicine Unlocked, 24, 100557. https://doi.org/10.1016/j.imu.2021.100557.

[22] Liu, G., Liao, Y., Wang, F., Zhang, B., Zhang, L., Liang, X., Wan, X., Li, S., Li, Z., Zhang, S., & Cui, S. (2021). Medical-VLBERT: Medical visual language Bert for COVID-19 CT report generation with Alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(9), 3786–3797. https://doi.org/10.1109/tnnls.2021.3099165.