CSE 572: Data Mining
Final Project Literature Review

Project Title: Medical Image Captioning and VQA using Multitasking for Chest X-Rays

Team members:

Full name	ASU ID
Amey Bhilegaonkar	1225368924
Disha Agarwal	1225441776
Ninad Nale	1225710226
Janaki Venkata Ramachandra Sai Nayani	1225418207

Step 1: Summary of relevant work

[1] Xu, J., Huang, J., Wang, C., Zhang, C., & Wang, J. (2021). Automatic ultrasound image report generation with adaptive multimodal attention mechanism. Artificial Intelligence in Medicine, 114, 102051. DOI: 10.1016/j.artmed.2020.102051.

Brief Summary

- Proposed a novel deep learning model that utilizes a multi-modal attention mechanism to generate radiology reports from ultrasound images
- The proposed adaptive attention mechanism learns to focus on different regions of the image as the report is generated, resulting in better report quality compared to baseline models
- Spatial attention and semantic attention models are proposed which utilize the output of the CNN model. This is helpful for locating objects in the partial regions of the image along with captioning

Strengths

- The proposed multi-modal attention model uses both visual and semantic features of the image, which might be useful for generating more accurate captions of the x-rays
- Spatial attention can improve image captioning, but it is not able to comprehensively describe ultrasound images, so a multi-label classification network is introduced to predict important localized semantic features to improve accuracy

Limitations

- Implementing this mechanism can be computationally complex, which may make it difficult to train and deploy the model on low-resource devices or in resource-constrained environments
- The model may require a large amount of annotated data to train the model, which can be challenging to obtain

[2] Yang, S., Niu, J., Wu, J., Liu, X. (2020). Automatic Medical Image Report Generation with Multi-view and Multi-modal Attention Mechanism. In: Qiu, M. (eds) Algorithms and Architectures for Parallel Processing. ICA3PP 2020. Lecture Notes in Computer Science(), vol 12454. Springer, Cham. https://doi.org/10.1007/978-3-030-60248-2_48

Brief Summary

- Proposed a novel attention-driven multimodal feature fusion model for medical image captioning, which combines both visual and textual features for generating captions
- The approach consists of a multi-view and multi-modal framework for generating a semantic-coherent medical report. Three main components of the framework: a multi-view image encoder, a medical concepts extractor, and a two-layer LSTM-based multi-modal attention generator for report generation
- The model is evaluated on the IU X-Ray public dataset of chest X-ray images. It outperforms several baseline models in terms of caption quality metrics

Strengths

- Injection of the semantic concepts which are the output of the concepts extractor into a two-layer (Attention & language) LSTM-based multimodal decoder generates more diverse semantically coherent reports
- Multi-view and multi-modal attention model performed significantly well compared to baseline models like CNN-RNN, Soft-Att, CoAtt, HRGR, KERP

Limitations

- Evaluation is done on a small public dataset of chest X-ray images, which may limit its generalizability to other datasets. Difficult to find more datasets that consist of frontal and lateral views, which might be a blocker
- Detailed analysis of the failure cases of the proposed model is not provided, which inhibits understanding of the limitations of this model and the potential for improvement

[3] Zeng, X., Wen, L., Xu, Y., & Ji, C. (2020). Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Computer Methods and Programs in Biomedicine, 197, 105700. https://doi.org/10.1016/j.cmpb.2020.105700

Brief Summary

- Proposes a new neural network called Semantic Fusion Network which includes a lesion area detection model and a diagnostic generation model
- The lesion area detection algorithm uses Faster RCNN to detect the lesion area and automatically separates that from the background and the lesion Area is used to get the encoding vector and pathological information
- In order to generate the words in the report we use an LSTM which uses the visual information to generate grammatical information and pathological information is mapped as semantic information. We use a sentinel gate to fuse semantic and grammatical information

Strengths

- Identifies the lesion area where there is abnormality
- It uses pathological information to generate more accurate reports
- It has higher accuracy compared to other models like Attention, NIC, m-RNN, etc.

Limitations

- The model misclassified some of the diseases which are similar to each other
- Pathological information of the generated report is extracted by keyword matching which leads to errors between generated and ground-truth information and it cannot be used to optimize parameters
- Some of the words in the diagnostic report may be wrong while generating long reports

[4] Yin, C., Zhou, Y., Zhang, X., Zhang, Y., Xu, D., & Wang, Y. (2019). Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In 2019 IEEE International Conference on Data Mining (ICDM) (pp. 728-737). DOI: 10.1109/ICDM.2019.00086

Brief Summary

- Proposes a new model to detect the abnormalities simultaneously and generates long diagnostic reports with paragraphs
- Uses a DenseNet CNN with a global label pooling layer for the multi-label classification problem to detect all the abnormalities
- The output of the DenseNet is fed to a Hierarchical RNN (with 2 LSTMs one with attention and one without attention) to generate a long medical annotation

Strengths

- Uses a Global Label Pooling layer instead of a global feature pooling layer in the CNN phase to make the model detect the abnormalities in the local part of the image instead of the global part
- Uses a Hierarchical RNN that has 2 LSTMs, one for the generation of topic vectors (sentence RNN) and one for the generation of detailed sentences based on topic vectors(Word RNN)
- Uses Attention mechanism in the sentence RNN to improve the quality of the focus particular abnormality

Limitations

- No labeling of the part where the abnormality takes place in the image
- Some models outperform the model on the CIDEr metric
- Does not use Reinforcement Learning, gets outperformed by models that use reinforcement learning

[5] M. Gu, X. Huang and Y. Fang, "Automatic Generation of Pulmonary Radiology Reports with Semantic Tags," 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), Jinan, China, 2019, pp. 162-167, DOI: 10.1109/ICAIT.2019.8935910.

Brief summary

- Proposes a new method for generating pulmonary radiology reports using an SRN-based attention mechanism and binary classifier with Semantic tags demonstrating the effectiveness of the method on a dataset of 4,000 radiology reports
- Experiments show that reports generated by semantic tags are of higher quality than reports generated by image features
- Combining the binary classifier with the model improves the accuracy of the generated reports

Strengths

- Provides a new method for automating the generation of radiology reports that can improve accuracy and efficiency
- Utilizes semantic tags to organize and structure the reports, which can be useful for further analysis
- The use of a binary classifier helps to increase the accuracy of the generated reports

Limitations

- Limited validation of the method on other datasets, which may have different characteristics
- The quality of the underlying data may determine how successful the semantic tags are and whether they are transferable to other domains
- The study only tested the model on a Chinese chest X-ray dataset, and it's unclear if it would perform similarly on other datasets from different populations

[6] R. Ambati and C. Reddy Dudyala, "A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering," 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 2018, pp. 1-6, DOI: 10.1109/INDICON45594.2018.8987108.

Brief summary

- Proposed a sequence-to-sequence model approach for visual question answering in the medical domain
- Uses pre-trained VGG-19 image feature extractor and GRU for encode-decoder layers, encoder encodes the question and decoder takes both question and image encoding to generate answer
- Achieved competitive results compared to other state-of-the-art models using ImageCLEF 2018 dataset

Strengths

- Proposed approach outperformed several other baseline models
- Implemented with a pre-trained image feature extractor to reduce training time

- The approach can be extended to other image-based question-answering tasks

Limitations

- Only evaluated on a single dataset(ImageCLEF 2018), so generalizability to other datasets is unclear
- Does not explicitly address issues related to bias in the dataset or the potential for models to perpetuate or amplify that bias
- More fine-tuning is required to improve performance, particularly for long and complex questions

[7] Hyeryun Park, Sanghoon Park, and Jeongyeon Kim. 2021. Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation. IEEE Access 9 (2021), 150560–150568. DOI:https://doi.org/10.1109/ACCESS.2021.3124564.

Brief Summary

- Uses both patient and normal images and then takes feature differences as input to a decoder or transformer model for caption generation
- Compares feature difference techniques subtracting feature, element-wise dot product, and average pooling vs global average pooling
- For generating image features Resnet is used as an encoder, for the decoder part, 2 models are used Hierarchical LSTM with co-attention and transformer(6 transformer decoder blocks, the encoded input is k,v to the multi-head attention of decoder) models

Strengths

- Identifies that element-wise dot product and average pooling results in better results across most of the models
- Establishes that transformer(simple transformer) models perform better than Hierarchical LSTM, also for high dimensional transformer models as X-transformer global pooling results in better Bleu score
- Verifies the methods on IUC as well as the MIMIC-CXR dataset and also on X-LAN and X-transformer

Limitations

- Does not explain abnormal results - Like for the IUC and MIMIC CXR datasets on X-LAN and X-Transformer have a better Rouge-L score for global average pooling and subtraction
- Could have further tested on CNNs such as Densenet121 or Faster RCNN

[8] Anderson, P. et al. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '18). IEEE, Salt Lake City, UT, USA, 6077-6086. DOI:https://doi.org/10.1109/CVPR.2018.00636.

Brief Summary

- Proposes 3 new methods to take into consideration image regions that affect the answer/caption. Uses bottom-up (what humans do is look at questions and then at the image) and top-down approach
- Bottom-up features are generated by Faster-RCNN with Resnet 101 to output regions having higher class detection probability
- The top-down attention model for image captioning uses CNN to get image regions/features(not bottom-up features), for captioning these are fed to an LSTM layer which acts as top-down attention and computes weights for each feature, the output of this LSTM and weighted(weights are calculated by the below layer) image features are used for the final layer
- Top-down attention model of VQA the model concatenates the CNN features with the embedding of the question(output of GRU layer) and generates bottom-up attention features which are fed to subsequent layers, which generate multi label output(the VQA task is converted to multilabel classification)

Strengths

- Improved all the metrics for the general domain dataset when it was written(2018)
- Incorporates bottom-up and top-down approaches, making the results more interpretable, the answer can be mapped to ROI's having high attention in the LSTM model
- A novel approach that can be adapted to other domains as well and sparks further research, is referenced by [3]

Limitations

- We can use bottom-up features for future research

Step 2: Organization of relevant work

Biomedical image captioning and VQA has got attention from researchers in both the language and vision domain, most of the papers use a CNN to generate image features and attention-based LSTM framework([1], [2], [3], [4], [7], [8]) for generating the output caption/answer. Some papers have converted the problem to a multilabel retrieval system and use fully connected layers or a softmax layer on top of the LSTMs like in [8] and [3] respectively. Others [4], [8] use 2 LSTM layers in which the 1st is called Sentence LSTM(generate a topic for a sentence), and the other is called Word LSTM which gives the final output.

Further, some papers have used multiple image inputs, [2] uses both lateral and anterior x-ray image features as input to the LSTM network and [7] uses the difference of image features from a normal image and patient image to generate the final feature vector. Both papers use the IUC dataset, [4] also uses it and it is a common dataset used for x-ray image captioning. Imageclef is another common dataset that is used in [4] and [6].

[1], [2], [3], [4] and [5] use semantic tags or labels as input features for better accuracy- [2], [4] combine it for input, [1] use attention on both image and semantic tags and uses that as input to LSTM layer. The

tag space is usually a high-frequency word in training set captions/answers or uses labels/classification results from CNN. Moreover, to further have more relevant image features [3] and [8] use Faster- RCNN to generate the Lesion/ ROI most relevant to the final output.