END-TERM REPORT

on

Explainable Medical Diagnosis Using Deep Learning

submitted in requirement for the course

B. TECH. PROJECT (CSN - 400A)

OF BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING

by

Aarush Gupta (16114002)

Aniruddha Mahapatra (16115021)

Dakshit Agrawal (16114022)

Under the Supervision of

Prof. Balasubramanian Raman (IIT Roorkee)

Prof. Marco Pedersoli and Prof. Jose Dolz (ETS Montreal)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

ROORKEE - 247667 (INDIA)

Contents

1 Introduction **4**

2 Overview of the Work Done **5**

3 Literature Review of Medical Diagnosis Generation using Deep Learning **5**

3.1 A Survey on Biomedical Image Captioning [1] 5

3.2 MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network [2] 6

3.2.1 Image Model 6

3.2.2 Language and AAS model 6

3.3 Towards Automatic Report Generation in Spine Radiology using Weakly Supervised Framework [3] 7

3.3.1 Recurrent GAN 7

3.3.2 Prior Knowledge based captioning 8

3.4 TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays [4] 8

3.4.1 Text-Image Embedding Network 8

3.5 CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison 9

3.5.1 Ignoring : 9

3.5.2 Binary Mapping : 9

3.5.3 Self-Training : 9

3.5.4 3-Class Classification: 9

3.6 On the Automatic Generation of Medical Imaging Reports [5] 10

3.7 Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation [6] 10

3.7.1 Image Encoder 11

3.7.2 Sentence Decoder 11

3.7.3 Retrieval Policy Module 11

3.7.4 Generation Module 11

3.8 Unsupervised Multimodal Representation Learning across Medical Images and Reports [**?**] 12

3.9 Clinically Accurate Chest X-Ray Report Generation [7] 12

3.10 Reinforced Transformer for Medical Image Captioning [8] 13

4 Literature Review for Novel Ideas in Medical Diagnosis Generation **15**
4.1 Stacked Semantics-Guided Attention Model for Fine-Grained Zero-Shot Learning [9] 15
4.2 A Dual Attention Network with Semantic Embedding for Few-shot Learning [10] and Dual Attention Network for Scene Segmentation [11] 16

5 Baselines **16**
5.1 Show and Tell: A Neural Image Caption Generator [12] 16
5.2 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [13] 17
5.3 On the Automatic Generation of Medical Imaging Reports [5] 18

6 Datasets **19**
6.1 IU X-Ray [14] 19
6.2 MIMIC-CXR [15] 19

7 Evaluation Metrics **20**
7.1 BLEU [16] 20
7.2 ROGUE [17] 20
7.3 CIDEr [18] 21

8 Experimental Setup **21**
8.1 Tools Used 21
8.1.1 Hardware 21
8.1.2 Software 22
8.2 Model Specifications 22
8.3 Data Augmentation 23

9 Results **23**

10 Future Plan **24**

11 Conclusion **25**

Abstract

Deep Learning models for report-generation, and medical diagnosis in general, do not fully explain the reason behind their diagnosis, making it difficult for doctors to accept deep learning solutions for fear of a fatal mistake, even though the diagnosis done are much better than humans. Providing better feedback for the reasons behind a certain diagnosis by a deep learning model will help assure doctors of its reliability, helping in faster detection of life-threatening diseases. We work towards creating a reliable and accurate deep learning model for medical diagnosis of certain diseases. We have implemented and tested deep learning models on the IU-XRAY [14] dataset to have some strong baselines, and based on the literature review we carried out this semester, we will work towards implementing novel ideas in the next semester. The benefits of a successful project are tremendous, with a potential to improve the medical field on both academic and industrial fronts.

1 Introduction

Medical image diagnosis is a prevalent method to detect a wide spectrum of diseases in the human body. It involves a careful inspection of the medical images by well-trained physicians, often followed by elaborate report-writing and corresponding inference. However, such methods can be error-prone if done by inexperienced physicians, or even time-consuming and tedious for experienced physicians. Many regions in the world do not have access to world class health facilities, hindering the chances of people living there to have an accurate and early diagnosis of possible medical diseases.

Artificial Intelligence may be used to complement physicians and doctors in diagnosing medical images of patients. With the abundance of patient data available, an AI model can learn to infer which conditions are symptoms of a particular disease. In places where healthcare facilities are poor, these AI machines can help doctors spot certain anomalies in the image. Such AI models are also very quick, being able to diagnose 100-150 images per second, with no dip in performance over time. Researchers have used and evaluated AI models on a variety of tasks, showing that they can diagnose medical images better than experienced doctors, and in much quicker time. Various other methods based on the application of neural networks have been tested with varying levels of accuracy on the task of medical image diagnosis.

Despite the advantages offered by such AI models, they have not been integrated in hospitals, or put to use for medical diagnosis. A critical field like healthcare leaves no room for mistakes, since even a small one could be fatal for the patient. A major hurdle in integrating the said AI models is the lack of trust that doctors have in them. The models must account for the predictions, making it difficult for doctors to accept such solutions in practice. Providing better feedback, including the uncertainty in the predictions and reasoning behind a certain diagnosis would help assure doctors of the reliability of such models.

A related task is that of captioning real-world images, known as visual image captioning. Our task is different from visual image captioning in that it involves generating long and topic-coherent reports. These reports also cover specific terminology/phrases depending on the medical task at hand. Hence, the models used for visual image captioning might not be relevant to our task. In our project, we tackle the lack of explainability in current AI medical image diagnosis methods.

2 Overview of the Work Done

This semester, we have mainly focused on the following:

- Carrying out a comprehensive literature review to find gaps in the existing research of explainable deep learning models for medical image reports (Section 3), as well as literature review to fill these gaps with novel ideas (Section 4).
- Implementing strong baselines to compare our novel ideas with. These baselines are the current state-of-the-art AI models, which we have coded and trained from scratch.

A detailed description of each baseline is provided in Section 5. The datasets used for training these baselines are elaborated in Section 6. The evaluation metrics are explained in Section 7. The environmental setups for these baselines are listed in Section 8, followed by an analysis of the performance of these baselines on the mentioned datasets and evaluation metrics in Section 9.

3 Literature Review of Medical Diagnosis Generation using Deep Learning

We did a literature review of some successful deep learning methods used for medical image diagnosis to determine plausible baselines to implement. It also helped us find gaps in the research done until now, which will be our areas of focus.

3.1 A Survey on Biomedical Image Captioning [1]

Automatic image captioning using deep learning models in the field of radiology and biomedical images can be quite useful for radiologists and doctors. The survey is a first of its kind on biomedical image captioning models, mentioning datasets, evaluation measures for analysing the results and the state-of-the-art models for medical image captioning.

3.2 MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network [2]

One of the shortcomings of machine-learning diagnostic models is the inability to provide the reason for a specific prediction or diagnosis. MDNet tries to overcome this shortcoming by establishing a direct multimodal mapping between the image and diagnostic medical reports. MDNet primarily focuses on providing better diagnostic reports on bladder cancer images (BCIDR dataset). This model reads the medical images, generates a diagnostic report from the image and can also retrieve images by symptom descriptions, and visualize attention-maps on specific parts of the image based on the medical report generated, to provide justifications of the network diagnosis process. The architecture is based on end-to-end training of an image model and language model. The image model is used for enhancing multi-scale features ensembles, while the language model, equipped with a better attention mechanism, aims to tries to learn a direct mapping between sentence words and image pixels from training data.

3.2.1 Image Model

The image models particularly use Resnets for extraction embeddings from images. The skip connections in Resnet allows the gradient to pass to the next block without passing through convolution layers, thus alleviating gradient vanishing. This can be viewed as 2^n different shallow models in Resnets, where n is the number of convolution blocks in the model, i.e., gradient from a convolution block can either pass through the preceding block or pass directly through skip connection, creating an ensemble model. However, one shortcoming of this model is that the weights to the output of individual convolution blocks remain undetermined. The authors of this paper provide and method to decouple the network and assign individual weights to each block, called by the authors as ensemble-connection. All these decoupled blocks are coupled in the last convolutional layer.

3.2.2 Language and AAS model

The language model is based on LSTM to learn diagnostic reports my maximizing the joint probability of sentences in the report. The LSTM uses a soft attention context vector z_t which dynam-

ically computes an attention weight vector from image based on the description. The attention mechanism, however, produces attention heat maps that highlight the entire image smoothly. To tackle the problem the authors use an extra level of supervision on attention named auxiliary sharpening model (AAS), that takes advantage of localization property of global average pooling layers of Resnet to localize attention maps in the image more prominently.

3.3 Towards Automatic Report Generation in Spine Radiology using Weakly Supervised Framework [3]

Supervised learning for generating diagnostic reports for medical images is difficult to achieve, especially in real-life scenarios. One of the reasons being a large amount of dataset that would be used to train the image-language models for supervised diagnostic report generation. An extensive corpus of radiologist reports is difficult to obtain unlike the large natural image captioning MSCOCO dataset. Another problem is due to the radiologist bias. Different radiologists write reports in different styles and structures and it is difficult to learn these structures from small amounts of data. This paper tries to alleviate this problem by adopting a semi-supervised approach to radiologist report generation in spine radiology using lumbar spine MRIs. It uses on object-level annotations of image bounding box and labels instead of full radiologist report. The authors use RCGAN for producing segmentation and classification information from spinal images and strong prior knowledge-based program for labeling and diagnostic report synthesis.

3.3.1 Recurrent GAN

The generative model produces a pixel-level segmentation of the entire image into 7 classes using the ACAE module and RLSTM based RNN model. The ACAE convolutional blocks help to increase the receptive field of the kernels by inserting zeros between each convolution kernel thus also minimizing loss which performing pooling. This helps to address the fine-grained details and variability in spinal images. The RLSTM based RNN model effectively memorizes spatial correlations among image locations that are quite useful, as there can be a high probability of abnormality in regions where neighboring regions are abnormal. The discriminative model tries to efficiently discriminate between original images and fake images generated by the generator.

3.3.2 Prior Knowledge based captioning

The output segmentation map produced by the RCGAN is used to produce 3 dictionary corresponding to location and normality of 3 spinal cord structures. The dictionaries are used to produce diagnostic reports.

3.4 TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays [4]

Some of the difficulty associated with supervised learning in diagnostic report generation for medical images is a shortage of large scale learnable datasets and radiologist bias in the report as was mentioned earlier. One more difficulty is the lack of techniques that can mimic high-level human reasoning obtained from years of radiologist knowledge and experience. This paper shows that free-text radiologist reports can be used as a priori knowledge for tackling these three problems. It uses a text and image embedding network along with multilevel attention mechanisms similar to MDNet to achieve this task. But unlike MDNet it works on Chest X-rays (OpenI dataset). A radiologist report consists typically of more vocabulary than just the keywords corresponding to different ailments and symptoms. They are just used as text embeddings for the model. The reports mainly consist of a finding section for different abnormalities in organs and inference section that lists diagnosis for the ailments.

3.4.1 Text-Image Embedding Network

The authors use the attention mechanism in RNN for word prediction with the initial context being the output of image embedding generated by Resnet50. In each state of RNN, the image embedding is modified by multiplying with different attention weights. The attention map in each hidden state is passed through the global average pooling layer to generate a combined attention map for all observations.

3.5 CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison

This paper presents a large scale chest X-ray dataset, CheXpert, consisting of Chest X-ray images along with radiologist reports for a large number of patients. The authors design a labeler that detects the presence of 14 observations in the radiologist reports along with uncertainty labels. The authors also investigate different approaches to use uncertainty labels in convolutional networks for obtaining the probabilities corresponding to different observations (aliments) from frontal and lateral X-rays. Each image has labels for 14 different observations. Each observation can have one of the 2 labels, namely, 1, 0, or u (uncertain). For multiple views of the same image, the output predictions are mean of all the predictions. To handle uncertainty the authors use 4 different types of approaches:

3.5.1 Ignoring :

The model takes binary cross-entropy loss for 14 observations and simply ignores loss corresponding to uncertain labels in the input image.

3.5.2 Binary Mapping :

The uncertain labels are replaced entirely by either zeros or ones.

3.5.3 Self-Training :

The uncertain labels are treated as unlabeled examples, thus this becomes a problem of semi-supervised multi-class classification. The model is initially trained using the Ignoring method. Then the output of the model for uncertain labels is treated as their labels for successive training.

3.5.4 3-Class Classification:

This method treats uncertainty labels as separate labels from 0 or 1 labels. Loss is taken as the average of multi-class cross-entropy rather than the mean of binary cross-entropy over observations.

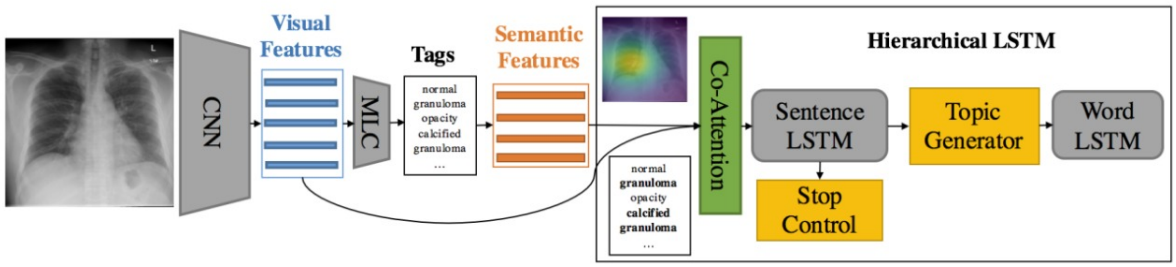

Figure 1: Model Proposed by Hsu et. al. in [5]

3.6 On the Automatic Generation of Medical Imaging Reports [5]

There are 3 more difficulties associated with medical image captioning in addition to the above-mentioned challenges. First, the abnormal regions are difficult to detect and locate in a medical image. Second, the diagnostic report contains heterogeneous information including the location of ailment, tag, and lebels. Third, the medical diagnostic reports are generally quite long, containing compound sentences. To cope with these 3 problems. This paper tackles the task by making some modifications to the CNN-RNN framework commonly used for captioning tasks. It first extracts the visual features from X-Ray images using a CNN. The visual features are used to extract some tags (using a multi-label classification network) in a setting similar to predicting multiple labels for an input. The tags are further processed to obtain vectors. The vectors are fed into a co-attention module along with the visual features. The co-attention module simultaneously attends to the visual and semantic features. Fig. 1 describes the entire process of obtaining tags ans vector for generating diagnostic report from image.

3.7 Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation [6]

Most of the medical diagnostic reports are quite normal without anomaly. Hence a typical generative model production sentences from images will likely focus more on the quality of sentences (that they are natural-looking) rather than capturing the minor anomalies, present in a small amount in training dataset. The authors use a model that consists of the following parts trained end-to-end:

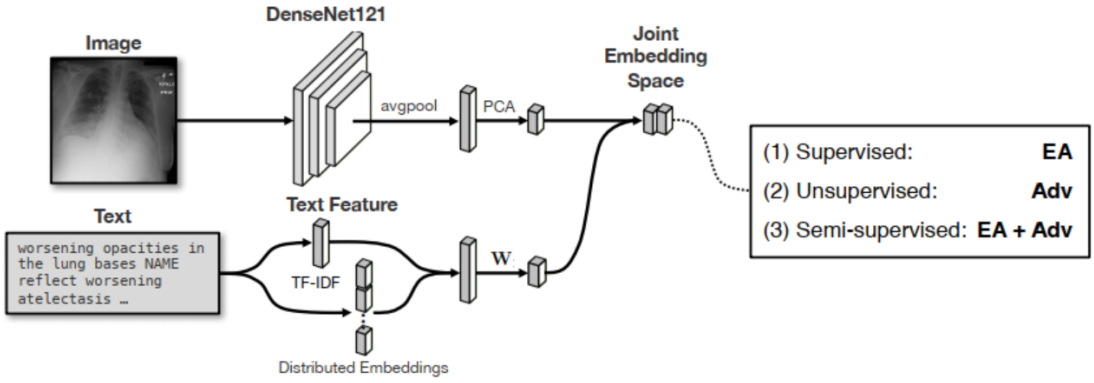

Figure 2: Model Proposed by Hsu et. al. in [**?**]

3.7.1 Image Encoder

The output of the last convolutional layer is used as a context vector for later modules (image embedding). In this paper, the authors use a pretrained VGG-19 or Densenet.

3.7.2 Sentence Decoder

The sentence decoder is a stacked RNN equipped with attention mechanisms for facilitation long term dependencies. The RNN produces a series of topics for the sentences.

3.7.3 Retrieval Policy Module

Since most of the medical diagnostic reports are formal and contain similar sentences that generally represent normal conditions, a database of all the sentences from known reports is maintained. This module predicts for each topic vector if a new sentence should be generated from it (probability near 0) or a template sentence can be taken as it is from the database (probability near 1). If it is the former case, the generation module is used to generate a new sentence, else the most appropriate template sentence is taken from the database.

3.7.4 Generation Module

The generation module for each topic vector generates a new sentence also taken the image context vector as the initial hidden state of the RNN.

3.8 Unsupervised Multimodal Representation Learning across Medical Images and Reports [?]

This paper exploits joint embeddings space of the visual and semantic features to allow for cross-domain retrieval and conditional report generation to provide accurate results. The main contributions of this work are as follows:

- Establish evaluation metrics and baselines for embedding-based report generation via retrieval and distance metrics in the embedding space.
- Document the relation between supervision level and representation quality in joint embedding spaces.
- Document the effect of using different sections of the report on the learned representations.

The text is reduced to embeddings by four methods: TF-IDF over bi-grams, Glove word embeddings, sentence embeddings and paragraph embeddings. The visual features are obtained from the pre-final layer of a DenseNet-121 model and are further reduced to obtain 64 dimensional vectors. The paper reports the results for five different types of alignments: linear transformations, adversarial domain adaptation, procrustus refinement, semi-supervised alignment and orthogonal regularization. Comparison between different methods are provided on the MIMIC-CXR dataset [15].

3.9 Clinically Accurate Chest X-Ray Report Generation [7]

This work states that majority of the work done has been concerned with producing real-looking and coherent reports. But these lack in clinical accuracy and are seemingly unable to capture the template-like structure inherent in medical reports. To counter this, the authors propose a Clinically Coherent Reward derived from metrics described in the CheXpert paper by Irvin et. al. [19]. The results are an improvement on NLG metrics such as ROGUE score, CIDEr and BLUE scores, as well as the clinical efficacy metrics than other models. This work also uses reinforcement learning to train the model as the proposed optimization metrics are non-differentiable, but can be optimized through policy gradient learning.

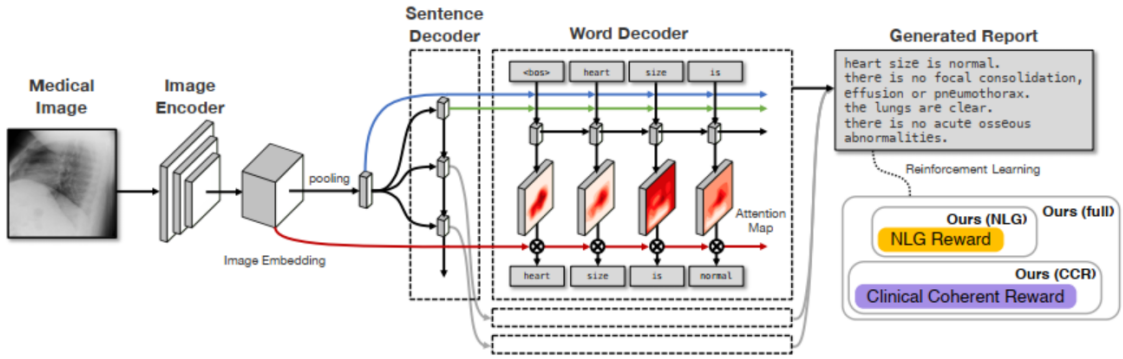

Figure 3: Model Proposed by G. Liu et. al. in [7]

Fig. 3 describes the methodology adopted by this work. The image is first passed through an image decoder (with the last global pooling layer removed). Further linear operations such as averaging reduces the feature map obtained into a single vector representation. This representation is fed into an LSTM network to predict the topics for each sentence (similar to [6]). Each topic vector is fed into another LSTM to predict words of the corresponding sentence one by one. In the word decoder, the model employs a visual sentinel vector [20] to augment the results by giving the model a chance to *look away* at the entire image representation as needed. To train the model, the authors transform the NLG metrics in the form of a loss function. This loss function is further augmented by using a clinically coherent reward (CCR). CCR uses the sparsity of diseases in the labels to provide proportionate rewards to the model when it makes the correct predictions.

3.10 Reinforced Transformer for Medical Image Captioning [8]

This paper recognized and tries to alleviate three major problems found in previous works:

- The visual encoders found in previous methods rely on top-down approaches to find suitable representations of the input images.
- Almost every previous work relies on recurrent architectures to predict the sentences from visual features. Recurrent architectures are unable to fully utilize the computation resources available due to the inherent time depencence of predictions in them.

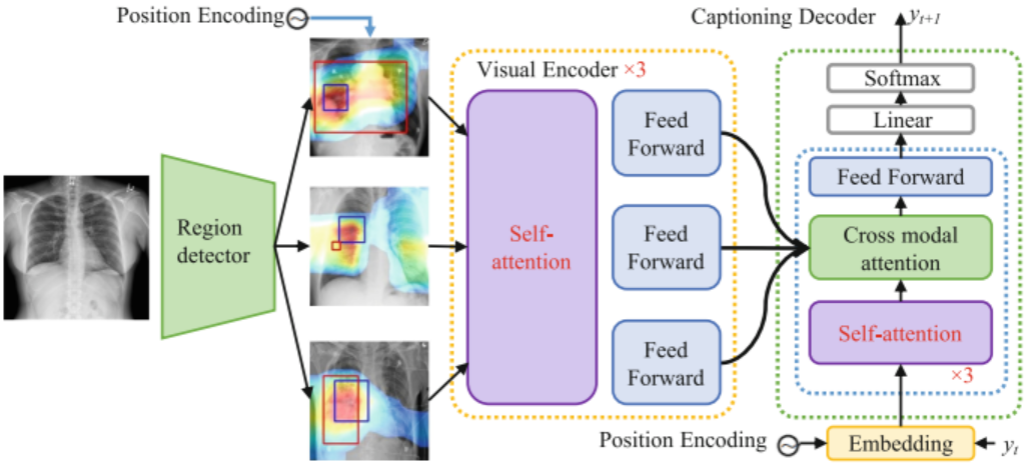

Figure 4: Model Proposed by Y. Xiong et. al. in [8]

- Continuous Maximum Likelihood Expectation is not able to account for the discrete task of language generation.

This work tries to incorporate bottom-up features from the input images to get better predictions, at the same time utilizing transformers proposed by Vaswani et. al. [8]. At the same time, the model proposed uses reinforcement learning to eliminate the exposure bias and the discrepancy between MLE and metrics used to evalute the predictions.

As shown in Fig. 4, the model first uses a Region detector (based on DenseNet [21]) to extract the relevant regions of the image. The DenseNet has been pretrained on the Chest X-Ray 14 dataset as for domain adaptation. These regions are then fed into a visual encoder (enclosed in the yellow box) to get the attended features of each region. This visual encoder acts as the top-down attention module, and is composed of 3 stacked multi-head self-attention layers and a fully connected layer.

These attended regions are then fed into a Captioning Decoder. This module consists of three parts: a self-attention module, a cross modal attention module and a simple feed-forward module. The decoder generates a separate sentence for each attended region from the visual encoder.

The whole model is trained using a semantic loss using the CIDEr score as a reward for word generated. The model uses the IU Chest X-Ray dataset for training and evaluation.

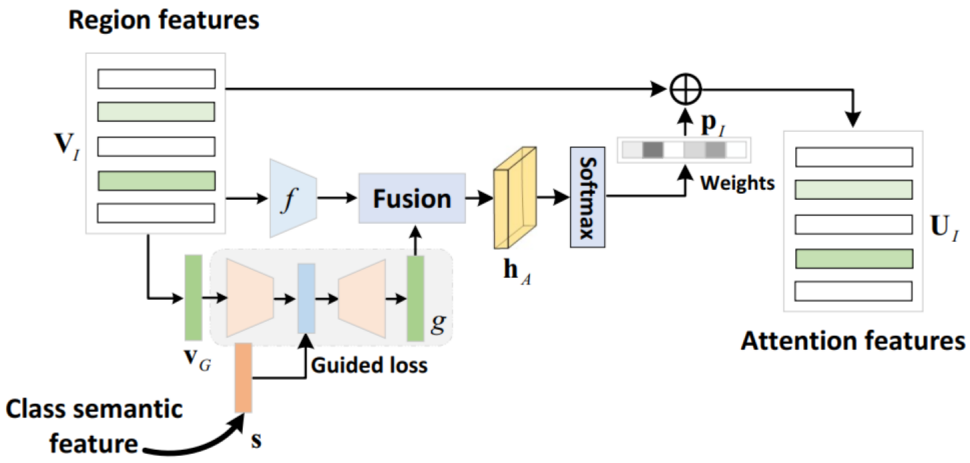

Figure 5: An illustration of the Semantics-Guided Attention (SGA) Layer

4 Literature Review for Novel Ideas in Medical Diagnosis Generation

We did a literature review of some successful techniques used in deep learning for other tasks, which have not yet been applied to the task of medical image diagnosis (existing gaps in current research). This review can help us determine potential directions to work and improve the current AI methods, both in performance and interpretability.

4.1 Stacked Semantics-Guided Attention Model for Fine-Grained Zero-Shot Learning [9]

This work proposes to use fine-grained visual features to get better discriminative information to distinguish classes. This approach is based on humans' nature to distinguish similar classes by looking at local regions, disposing off information from similar regions and locating the dissimilar ones for discrimination. To do so, the authors use an attention mechanism to focus on such regions. They utilize a stacked attention network guided by class semantic features for Zero-Shot Learning.

The main component of the proposed model that is relevant to the problem at hand is shown in Fig. 5. The attention map is extracted from two different networks in the following manner:

- A local embedding network projects the local visual regions into a latent space.

- Another semantic-guided network compresses all the local features to a single vector using a three-layer encoder-decoder framework. The output of the middle layer is forced to be close to the corresponding semantic feature.
- The reconstructed visual feature from the second step is then used to obtain an attention map on the local visual regions.

4.2 A Dual Attention Network with Semantic Embedding for Few-shot Learning [10] and Dual Attention Network for Scene Segmentation [11]

Both these papers approach different problems with the same idea: using two attention modules to attend to different representations viz. spatial and channel input, and the spatial and task-specific information.

- The first paper couples a meta-learning network to obtain spatial attention maps from the input image to produce a selective pooled visual vector. It further uses task attention to learn the importance of each training example for the given test image class.
- The second paper uses self-attention modules to separately attend to the positional and channel-wise visual information. These two attended representations are then summed element-wise and passed through a convolution layer to predict the segmentation maps.

5 Baselines

In this section, we talk about the deep learning models that we have implemented. We will be using these models as baselines to evaluate the performance boost given by our novel idea.

5.1 Show and Tell: A Neural Image Caption Generator [12]

This research paper introduced a very simple deep learning model to tackle the task of image captioning. It has served as baselines for most of the following research done in image captioning, and even medical report generation. It consists of a visual deep CNN to encode the image into characteristic feature vectors, which are then used by the language generating RNN to generate captions

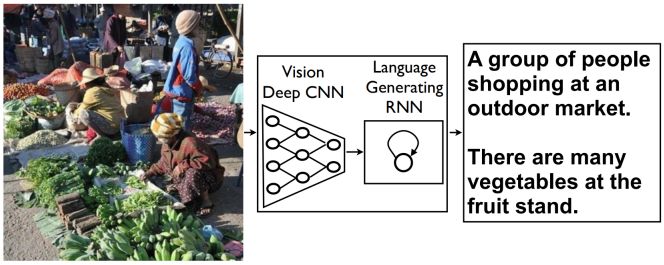

Figure 6: Proposed deep learning model in [12]

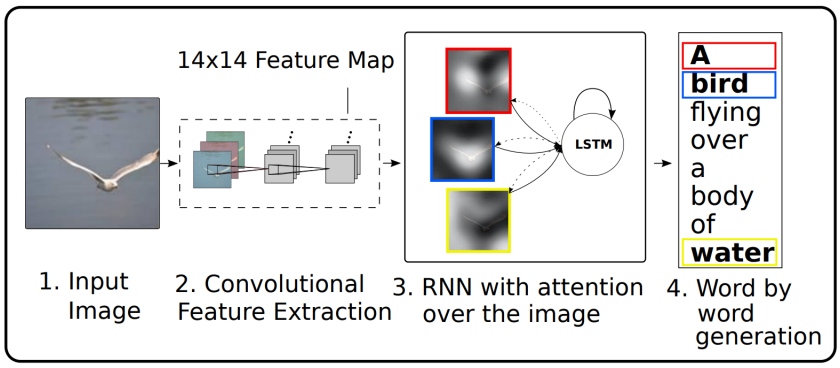

Figure 7: Proposed deep learning model in [13]

(see Fig. 6). The simplicity of this model means that there are many scopes for improvement, which have been explored by succeeding works. We implemented the exact same architecture to use as a baseline.

5.2 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [13]

This research paper built upon the technique proposed in [12]. They took cues from how humans describe a picture. Humans generally look at certain areas of the picture, rather than the whole picture, while describing the picture. This was termed as ”attention” by the authors, which they introduced into the deep learning model proposed earlier in [12].

The attention model works by taking a weighted sum of the image encoding vectors generated by the visual deep CNN. These weights are generated by taking a dot product of the current hidden state vector of the language generating RNN with each of the image encoding vectors. The dot product denotes similarity, with image encoding vectors more similar to the current hidden state

vector having higher weights. We implemented the exact same architecture to use as a baseline.

5.3 On the Automatic Generation of Medical Imaging Reports [5]

This paper introduced a strong baseline for medical image captioning. It still used a visual deep CNN to generate the image feature encoding vectors, but instead of a simple language generating RNN, a hierarchical LSTM, introduced by Krause et. al. [22], was used (see Fig. 1). A hierarchical LSTM had two seperate LSTMs, a sentence LSTM to generate the overall idea or topic of the sentence, and a word LSTM to generate the actual words for the sentence, based on the topic generated by the sentence LSTM. This structure helped in remembering information for longer times. Since a sentence contains around 10-12 words, the time steps for which information is needed to be remembered by the word LSTM is much shorter.

Apart from the above, the authors integrated two novel ideas into this pipeline:

- an auxiliary multi-label classifier loss to generate the tags (these tags denote diseases)
- a co-attention mechanism. This mechanism worked by taking the current hidden state vector of the sentence LSTM to calculate weights with both the visual feature encodings as well as the semantic feature encodings according to [13]. The weighted visual feature vector and semantic feature vector were then concatenated and passed through a linear layer to generate a co-attention vector, which is then used as input to the sentence LSTM.

For our implementation of the deep learning model proposed by [5], we did not have access to the tags. Hence, we have not used the auxiliary multi-label classification loss, nor does the co-attention model use the semantic feature encoding vectors of the tags, due to their unavailability. We use the weighted sum of the following losses to train the model:

- Cross-entropy loss of stop vectors (denotes to stop producing more topic vectors) (Loss 1)
- Cross-entropy loss of words predicted by the word LSTM (Loss 2)

Table 1: Dataset Statistics

Dataset	Training Instances	Test Instances	Total Instances
IU X-Ray [14]	6,674	756	7,430
MIMIC-CXR [15]	205,050	22,785	227,835

6 Datasets

We plan to evaluate our deep learning models on two datasets (see Table 1) of different sizes to have a comprehensive idea of the performance that our models achieve. Both of the datasets are used for training AI models to provide medical diagnosis for Chest X-Rays, highlighting potential diseases in the chest area.

6.1 IU X-Ray [14]

Demner-Fushman et al. (2015) [14] presented an approach for developing a collection of radiology examinations, which included narrative reports and images by radiologists. The authors suggested an accurate anonymization approach for textual radiology reports. The IU X-Ray dataset is publicly available if it is searched on the Open Access Biomedical Image Search Engine (OpenI).

The dataset contains 7,470 frontal and lateral chest X-rays and their corresponding radiology reports, labeled by experts. Each radiology report has four sections, but the AI models will predict only two of them. The **comparison** section contains previous information about the patient (e.g., preceding medical exams); the **indication** section contains symptoms (e.g., hypoxia) or reasons of examination (e.g., age); **findings** lists the radiology observations; and 'impression' outlines the final diagnosis. An AI system generally generates the 'findings' and 'impression' sections. Some researchers (Jing et al., 2018 [5]) have built AI systems that generate these sections together, i.e., concatenated.

6.2 MIMIC-CXR [15]

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 [15] is a very large dataset which is available on Physionet. Its access is limited, being granted only when a special ethics course is

completed by the user. The dataset consists of chest X-Rays in DICOM format with their corresponding radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. In order to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements, the dataset has been de-identified (specific patient names and sensitive information replaced by 'xxxx'). Protected Health Information (PHI) has been removed. The dataset is intended to support a wide body of research in medicine including image understanding, natural language processing, and decision support. Such large datasets help in reducing overfitting while training the AI models, which is useful in achieving better performance while using these models to generate diagnosis for medical images it has never seen before.

7 Evaluation Metrics

There are standard metrics for evaluating sentences generated by deep learning models, viz., BLEU [16], ROGUE [17] and CIDEr [18].

7.1 BLEU [16]

The Bilingual Evaluation Understudy (BLEU) [16] metric is the most popular evaluation metric for natural language generation. It is calculated using the following formula:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

It is basically a modified n-gram precision, which is the fraction of n-grams in the candidate text which are present in any of the reference texts, penalized by the brevity in candidate texts (BP).

7.2 ROGUE [17]

The Recall Oriented Understudy for Gisting Evaluation (ROGUE) [17] metric is based on recall, as is evident from the name. There are various ROGUE metrics introduced by the authors. We use

the ROGUE-L metric, which is based on longest common subsequence (LCS). Suppose A and B are candidate and reference summaries of lengths m and n respectively. Then, we have

$$P = \frac{LCS(A,B)}{m} \text{ and } R = \frac{LCS(A,B)}{n} \tag{2}$$

F is then calculated as the weighted harmonic mean of P and R, as

$$F = \frac{\left(1+b^2\right)RP}{R+b^2P} \tag{3}$$

7.3 CIDEr [18]

The Consensus-based Image Description Evaluation (CIDEr) [18] metric evaluates how well a candidate sentence c_i matches the consensus of a set of image descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$. The evaluation is done with the following formula:

$$\text{CIDEr}_n\left(c_i, S_i\right) = \frac{1}{m}\sum_j \frac{g^n\left(c_i\right) \cdot g^n\left(s_{ij}\right)}{\left\|g^n\left(c_i\right)\right\| \left\|g^n\left(s_{ij}\right)\right\|} \tag{4}$$

Here, the function g denotes taking the Term Frequency Inverse Document Frequency (TF-IDF) of the input. These CIDEr scores are taken for n-grams of length n.

8 Experimental Setup

8.1 Tools Used

8.1.1 Hardware

We had access to the following two systems, provided by the Institute Computer Center (ICC):

- DGX-1 server
  - Ubuntu 18.04
  - 8 NVIDIA Tesla V100 GPUS (each of 32 GB)
  - 512 GB RAM

- One high-end system in RS lab, ICC
  - Ubuntu 18.04
  - 1 NVIDIA Quadro P5000 (of 16 GB)
  - 80 GB RAM

8.1.2 Software

We used the following environment and libraries for coding and training the model:

- Python 3.5
- Pytorch 1.3.0
- Torchvision 0.4.1
- Matplotlib for visualization
- NLTK for primary BLEU score evaluation
- CoCoCaption for final evaluation

8.2 Model Specifications

We specify the hyperparameters for our best model, which is the baseline mentioned in Section 5.3:

- **Learning Rate for CNN Encoder:** 1e-5
- **Learning Rate for Hierarchical LSTM:** 5e-4
- **Input Dimension of Sentence LSTM:** 1024
- **Hidden State Dimension of Sentence LSTM:** 512
- **Input Dimension of Word LSTM:** 512
- **Hidden State Dimension of Word LSTM:** 512

Table 2: Model Results

Model	BLUE-1	BLUE-2	BLUE-3	BLUE-4	ROGUE	CIDEr
Simple	0.213	0.039	0.010	0.003	–	–
Attention	0.241	0.071	0.0202	0.00	–	–
Heir	0.492	0.364	0.279	0.219	0.534	1.556

- **Weight of Loss 1:** 1
- **Weight of Loss 2:** 1
- **Encoder Architecture:** VGG-19 without pretrained ImageNet weights
- **Batch size:** 64
- **Epochs:** 50

8.3 Data Augmentation

We mention the data augmentations we used to train the baselines:

- Resize the image to have length of one side equal to 224
- Random crop of 224x224
- Random horizontal flip

9 Results

The baselines that we have implemented in Section 5 serve as strong baselines which any new idea we propose should beat in performance. All the 3 deep learning baselines were evaluated on the IU X-Ray dataset, whose results are given in Table 9.

We observed that due to the very small size of the IU X-Ray dataset, the models easily overfit. This is another challenge that evaluating on two datasets of different sizes puts forward, that of finding a model which achieves a satisfiable level of performance even with few instances of data.

Our results were comparable in BLEU scores with the respective papers, but did not match in ROGUE or CIDEr scores, with ours being much better. This led us to find huge inconsistencies in the results mentioned in the papers. We have decided to show our novel idea's performance against the scores we are achieving, rather than the ones mentioned in the papers, since there may be a difference in the environmental setup, or the train-test split.

We were unable to evaluate the performance of the baselines on the MIMIC-CXR dataset due to the extremely large size (nearly 5 TB). We will have a look at how other people in the community have used this dataset to train their models, and use it in the same way.

10 Future Plan

The extensive literature survey has helped us in coming up with the following ideas, which we plan to experiment with in the future:

- Using loss functions derived from semantic metrics for training the models. This helps the model learn to produce clinically accurate and semantically more meaningful reports.
- Using stacked attention modules and dual attention modules to help the model provide better attention maps that would produce more accurate reports as well as provide better feedback on the predictions.
- Using transformers instead of traditional RNN architectures, as they have provided better results on a variety of language tasks.
- Learning a joint space to learn aligned visual and semantic features.

The MIMIC-CXR dataset has a total size of around 5 terabytes, which cannot be completely utilized for training given the resources we have. Therefore, we also plan to prune this dataset to a reasonable size which we could use to train models on. We also plan to explore more ideas while working on these ideas. An ambitious goal would be to submit a research paper to ECCV in February, in the case that the ideas we implement give us good results.

11 Conclusion

We took up a B. Tech. Project that would help us apply the skills we learned in the field of deep learning for a bigger cause. To work in that direction, we are working towards improving reliability and explanability of medical diagnosis provided by deep learning models. We have read relevant papers and also implemented and tested some strong baselines on datasets to get an idea of the work that will be needed to be done. We look forward to implementing a novel idea which fills one of the gaps found in our literature review, and evaluating how much better it performs against the current baselines we have implemented.

References

[1] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos, "A survey on biomedical image captioning," *CoRR*, vol. abs/1905.13302, 2019.

[2] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network," *CoRR*, vol. abs/1707.02485, 2017.

[3] Z. Han, B. Wei, S. Leung, J. Chung, and S. Li, "Towards automatic report generation in spine radiology using weakly supervised framework," in *MICCAI*, 2018.

[4] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," *CoRR*, vol. abs/1801.04334, 2018.

[5] B. Jing, P. Xie, and E. P. Xing, "On the automatic generation of medical imaging reports," *CoRR*, vol. abs/1711.08195, 2017.

[6] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *CoRR*, vol. abs/1805.08298, 2018.

[7] G. Liu, T. H. Hsu, M. B. A. McDermott, W. Boag, W. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," *CoRR*, vol. abs/1904.02633, 2019.

[8] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Machine Learning in Medical Imaging* (H.-I. Suk, M. Liu, P. Yan, and C. Lian, eds.), (Cham), pp. 673–680, Springer International Publishing, 2019.

[9] Y. Yu, Z. Ji, Y. Fu, J. Guo, Y. Pang, and Z. Zhang, "Stacked semantic-guided attention model for fine-grained zero-shot learning," *CoRR*, vol. abs/1805.08113, 2018.

[10] S. Yan, Z. Songyang, and X. He, "A dual attention network with semantic embedding for few-shot learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9079–9086, 07 2019.

[11] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *CoRR*, vol. abs/1809.02983, 2018.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CoRR*, vol. abs/1411.4555, 2014.

[13] 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," *CoRR*, vol. abs/1502.03044, 2015.

[14] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, pp. 304–310, 07 2015.

[15] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR: A large publicly available database of labeled chest radiographs," *CoRR*, vol. abs/1901.07042, 2019.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "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), pp. 311–318, Association for Computational Linguistics, July 2002.

[17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[18] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *CoRR*, vol. abs/1411.5726, 2014.

[19] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg,

R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *CoRR*, vol. abs/1901.07031, 2019.

[20] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via A visual sentinel for image captioning," *CoRR*, vol. abs/1612.01887, 2016.

[21] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[22] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," *CoRR*, vol. abs/1611.06607, 2016.