MONASH UNIVERSITY

MINOR THESIS

Medical Report Generation Using Hierarchical Human Knowledge Oriented Image Captioning

Author:
Yang WANG

Supervisor:
Dr. Xiaojun CHANG

June 12, 2020

MONASH University

Contents

Declaration

I declare that this thesis 'Medical Report Generation Using Hierarchical Human Knowledge Oriented Image Captioning' and the work presented in it are my own. When I quoted works from others, the source is always given. This thesis has not been submitted to any university for any degree.

Signed: 王泽

Date: 2020. 06. 11

Abstract

This research aims to present a weakly supervised medical report generation model that could outperform current traditional supervised image captioning models. Medical imaging could be described as involving the hierarchical distillation of knowledge and logical sentence construction. Medical imaging is a task that involves the reading or interpretation of medical images and the writing of medical reports. This differs from tasks such as summarization and image captioning. Summarization tends to be more broad and generalised in its language, avoiding the use of ”internal” field-specific vocabulary and standardised templates. In image captioning, a single sentence is usually desired. The traditional method of supervised image captioning has faced many challenges in accomplishing the task of adequately interpreting and evaluating medical imaging.

Acknowledgements

This year is a special year. Due to the COVID-19 virusI can not go to University to study. I would like to thank my supervisor, Dr. Xiaojun Chang, for supporting me all the way to the end. Also, my friend JingMing Zhao who always discussed and shared ideas during the maing of this research. My friends Oscar Elmahdy and Jun Guo who supported me during the writing of this thesis. And all the members of the Machine Vision Group who helped me during experimentation. Last but not least thank you to my dear parents. Thank you all.

1 Chapter I

1.1 Overview

In the last several years, the challenge for Computer Vision (CV) has changed. The major challenge has changed from single image classification to multi-category, multi-instance classification, and multi-object detection. The image captioning task is one of the most challenging tasks we face today. Vinyals, Toshev, Bengio, Erhan (2015) proposed a sequence-to-sequence single sentence image captioning task (55) which could extract features from images by a deep Convolution Neural Network (CNN) model, and then used a Recurrent Neural Network (RNN) model to generate descriptions. Using deep Convolutional Neural Network (CNN) models as the encoder and a Long-Short-Term Memory (LSTM) model as the decoder. This method has become a standard structure in the image captioning area.

Li, Liang, Hu, Xing (2018) reported three main problems in image captining taks, since Vinyals et al has researched extensive image captioning work, (34): First, the generation of long descriptions with multiple-sentences or paragraphs, with consistent logic and topics. Second, the ability to describe scenarios using the language and prioritisation of details that a human would make, based on field-specific vocabulary and knowledge. For example, describing the players and details of a sports game using terms and expressions in the way that a human would describe them. Third, appropriate ordering of sentences according to importance. For example, having more generic or summary information at the beginning of a text, such as which team won the game, the score, and then expanding into greater detail. These are some of the biggest challenges faced in the task of image captioning.

Medical report generation represented an applied use cases of image captioning. It would potentially support radiologists read and interpret medical images more effectively. In some cases, radiologists might be inexperienced or lack expertise. A structured radiology report which is generated by a machine could support them recognize some conditions or diseases in medical images. Furthermore, it could help bolster rural or other under resourced clinics by reducing the dependence on experts for the reading of images. Current image captioning models have not performed sufficiently well in the task of medical image interpretation and evaluation. This is in part due to some of the challenging characteristics of the task, such as:

- A biased dataset. ”Normal” (healthy) reports and images far outnumber the ”abnormal” (positive diagnosis) images. Some categories of images only have a few examples, which makes it difficult to learn about and model abnormalities. Moreover the abnormal, rarer types of images contain the most important elements to be identified.
- Lack of differentiation of images. Patient A’s X-ray images and patient B’s X-ray images probably display markers for two very different diseases, but visually the images appear similar to each other, with only a small but critical difference. This makes it hard for the model to distinguish between them.
- Inconsistency of description. Because different doctors have different ways of describing the same findings, even more so when it comes to ”abnormal” images, the same image would be described in different ways by different doctors. This creates difficulties in modeling and identifying patterns.

1.2 Research Background

1.2.1 Image captioning

Vinyals et al (2015) (55) used the last layer of a CNN to represent image features, then fed them into an LSTM decoder, in such a way, the model could generate a sentence based on the given image. Xu et al (2015) (68) introduced a new mechanism which is spatial-visual attention. It helped CNN to capture significant features. Recently, one research group proposed a multi-attention model, which used both local region image features and word features together to represent the superior pair information. (Yu, Fu, Mei, Rui, 2017) (73)

1.2.2 Computer-Aided Detection

In the medical image processing field, there are two sub-fields which have become the main research problem, which is Diagnosis (CADx) and Computer-Aided Detection (CADe).(Chartrand et al, 2017) (7). Lately, the models using Deep Learning techniques progressively outperform traditional models in the image processing field, such as Esteva et al (2017) proposed an automated classification of skin lesions (16), Ben-Cohen et al (2016) applied Deep Learning at liver lesion detection (3), and Zhang, Chen, Sapkota, Yang used it for pathological image finding detection (75). However,

Liu et al(2017) mentioned that these approaches are commonly designed for binary classifications, such as colon polyps, lymph nodes, or lung nodules which only identify one specifically disease or lesion (40). The ideal goal of Medical image captioning task is to classify multiple lesions or diseases as accurate as possible, and then generate a description as logical as possible.

The Co-attention model simultaneously uses both visual and semantic features to generate detailed chest X-ray images report based on a hierarchical generation framework (Jing, Xie, Xing, 2017). The results of Co-attention model has become the start-of-art in that time on on the IU chest X-ray dataset(13). However, it still has some limitations such as there are some words which were repeatedly generated by the model. This is due to the Co-attention model lack of contextual coherence when generating the report (27).

To solve the repetition problem, a multi-model recurrent generation model was adopted.(Xue et al ,2018). It would generate next sentence by all the previously sentences. In detail, where the succeeding sentence would be conditioned upon multi-modal inputs that include preceding sentences and image local features, more consistent reports could be generated (69). The problem of rarer image types in an unbalanced dataset was solved by inputting human generated templates (Li, Liang, Hu, Xing, 2018) (34). They adopted the use of both templates and a RNN model to generate reports, then used Reinforced Learning to set a reward to train prioritisation of which method was most suitable for different scenarios. Using this method they also achieved great results using the IU X-ray dataset. Knowledge Graph was applied to connected groupings, and then generating sentences (Li et al, 2019) (32). All of the previously mentioned researchers have so far used traditional supervised image captioning methods to solve the difficulties of medical report generation.

1.3 Research Aim

Rather than traditional supervised image captioning, we seek to use a different approach in order to overcome the main difficulties these methods have had with the task of interpreting and evaluating medical imagery. Our goal is to outperform the previously detailed methods. This thesis aim to provide an improved method for Medical Report Generation (MRG), specifically for chest X-ray reports.

1.4 Research Design

In this thesis, we introduced a weakly supervised method in Medical Report Generation(MRG) task. The model include two key components. First, we designed a corpus filter model which uses template embedding and clustering to select information from a raw medical report corpus. It helped us to greatly reduce the noise of chest X-ray reports and the variations in writing style under different radiologists.

Secondly, we designed a model which could generate a report based on a given X-ray image through the use of templates. We then assign two different reward generators to help model sentence generation based on the existing templates. A Hierarchical Concept Reward assisted model to generate reports that resemble real human written reports. Finally, we combine both approaches to get the final report.

1.5 Outcome

We present a weakly supervised medical report generation model with characteristics including:

- A hierarchical knowledge concept mechanism with a practical concept distillation and noise reduction method.
- Our proposed model does not rely on paired image and report, only requires an X-ray image set and a medical report corpus with several raw concepts.
- Experimental results indicated that our approach has an ability to generate quite promising medical reports and shows state of the art results on different report benchmarks.

1.6 Evaluation

Two main evaluation methods will be used in this thesis.

Automatic Metrics: There is no typical evaluation matrix design for this kind of task. Thus, in order to compare

with other research's result we use same evaluation method as them. There are several traditional automatic metrics has been used in language generation task. BLEU, METEOR, and ROUGE were design for language generation task. CIDEr was design for image captioning task

Expert input:Physician in related area will be invited to evaluate and analyse the accuracy of the final generated reports.

2 Chapter II

2.1 Literature review

2.2 Overview

This thesis aims to design an weakly supervised method which could applied on medical report generation. This chapter will present a structured and comprehensive literature review in both Radiology and Deep Learning related.

Firstly, this chapter will begin with a concise introduction to recent radiology and how they have developed the use of Deep Learning. There are essentially five main areas of Deep Learning as applied to the field of radiology. The most popular is Classification, and its problems are also the most relevant to this research topic. There are also the areas of Detection, Segmentation, Registration, and "Other", and the problems faced in these areas are less relevant to this research.

Following the review will be an introduction to the different datasets in the area of chest X-ray, and several works which are based on these datasets. Additionally, what the advantages and limitations are of each dataset will be covered.

Next, the review will introduce the language generation and image captioning methods of recent years which have influenced this research.

Finally, the review will introduce the previous research that has been done in the area of medical report generation, and summarise some of the common problems the researchers faced.

2.2.1 Radiology

Diagnostic radiology is the specific medical area which required radiologists diagnose patients condition through a non-invasive medical image. Through these detailed medical images, radiologist could identify the main risk of various diseases and other diagnoses, and in combination with a individual's medical condition, they could summarise and write findings and impressions in the radiology report. Nevertheless, various radiologist have their personal ways of structuring reports and their own idiosyncrasies of style. For example, given a patient that displays no abnormalities, one radiologist might write something along the lines of "no acute finding", while another probably record "no marker A present, no marker B present ...". A part of radiologist might prefer to drawn a conclusion as final sentence of report, while other prefer to claim patient's condition in the first sentence.

These variations on the delivery of the same core message can lead to potential confusion in even clinician-to-clinician communication. Thus, it is important to improve the quality of the report. A study showed radiology physicians rate the reports accompanying referrals written by other physicians an average of 8 and 7.61 respectively out of 10 for quality and clarity, although most of agree the high importance of clarity in medical reports (Clinger et al, 1988 and Schwartz et al, 2011). If there is a structured and general format that could be followed when radiologist are writing a report, the communication would be much better among clinical diagnostic radiology. A general structured radiology report should include several key components, such as impression, finding, patient's history and reason for admission (Kahn Jr et al, 2009) (28). The patient's history should contains the individual clinical record. For instance, surgical or allergy history. The detail information and description are usually written in the finding section. For example, positive and negative, location and size, severity and transparency. On the other hand, the impressions section generally summarized all the other sections including finding section, provided a brief overview of current image.

2.2.2 A.I. and Radiology

In past few years, Deep Learning have been employed extensively in medical image processing. Generally, there are five areas of Deep Learning that are relevant to medical image processing: Classification, Detection, Segmentation, Registration, and "Other". (Litjens et al, 2017) (38).

Classification
In this area, there are two different of classification. The first one is the binary classification, or it also called single image or exam classification. Normally, the input could be a singe image or multiple images, while the output is a single diagnostic variable. For instance, "has such condition" or "does not display markers for such condition". There are some researches have studied in this field. Esteva et al (2017) introduced an automated skin lesion classification model (16). Zhang, Chen, Sapkota, Yang (2017) proposed the detection of pathological-image findings.(75). Ben-Cohen et al (2016) came up with the liver lesion detection model. (3) Convolutional Neural Networks (CNNs) is commonly applied in this type of classification. Most of them use the pre-trained CNN which trained on some common image datasets (non-medical image datasets) such as DenseNet (22) and ResNet (21).

The second is lesion classification or object classification. Similar to the binary classification mentioned above, the input could be single or multi-image, but in terms of output, it would give multi diagnostic variable, not only one variable. It usually focuses on smaller regions of the medical image. In order to achieve this goal and have high accuracy, it requires more information than binary classification, usually lesion appearance and lesion location would be considered local information and global contextual information. However, this kind of data is difficult to convert to the format which traditional Deep Learning architectures accept. There are several researchers who have been studying this task. By using three CNNs to extract different scale of nodule patch, then combine the output of these three CNNs to be the final output (Shen et al ,2015).(48). A similar approach being researched is the use of multi-CNNs, every network would concentrate on different part of the image to identify skin lesions. (Kawahara and Hamarneh ,2016)(30). Also it could add 3D information as well, MRI is suitable in this case. The features extracted by 3D CNNs are used to test survival time of patients with brain tumors (high-grade Glioma)

Detection
Similar to classification, there are two kinds of detection, organ region detection and localised landmarks, respectively. Localised landmarks are usually used in 3D medical images, as we discussed above, 3D information is hard to use in CNN architecture. In order to find a way to parsing 3D data into neural network, there are several approaches. One example is combining 2D information together to get 3D information. The model take three independent sets of 2D MRI slices (one for each plane) extracted using regular CNNs as input to identified landmarks on the distal femur surface. The output of 3D landmark position is the intersection of the three 2D slices with the highest classification (Yang et al ,2015)(70). Another example is the use of a rectangular 3D bounding box to localize regions of interest (ROIs) around anatomical regions (descending aorta, aortic arch, and heart) (De Vos et al, 2016) (11)

Detected objects or lesions from medical images is the task of organ region detection. In detail, it includes two subtasks, identify small lesions, and localisation from a medical image. Cerebral Microbleeds could be detected accurately by using a 3D Convolutional Neural Network extract features from magnetic resonance (MR) images. (Dou et al, 2016) (15) Lesion localization has achieved impressive results by using weakly-supervised Deep Learning architectures. (Hwang and Kim, 2016)(23)

Segmentation
Deep Learning is commonly used in medical image segmentation tasks which could identity voxel sets of an object regardless outline or interior. Using Deep Learning segments image feature from surgical and biopsy tissue specimens is helpful to predict the invasion degree of the disease and provide the basis for the diagnosis and classification of the disease. Based on histopathological images, it could extract the most valuable image features. Until recent years, CNN is widely applied in histopathological images segmentation and microscopic image segmentation. A deep CNN model was applied to medical image segmentation using a sliding window. In this way, the bio-neural membrane was segmented in to an electron microscope image.(Ciresan et al, 2012) (10). From cardiac MRI data, if the model could segment the left ventricle, then calculating clinical indicators such as contraction rate and ventricular volume could be much easier which could help clinic surgery effectively. One research group designed a model for left ventricle appearance by using DBN learning feature, and segment the left ventricle into the echocardiographic image automatically using a supervised learning model.(Carneiro et al, 2018) (6). Besides this, another research group improved left ventricular segmentation

Chest X-ray dataset	Institution	language	images	reports
IU X-ray	Indiana Network for Patient Care	English	8,121	3,996
Chest X-ray	National Institutes of Health	English	108,948	0
CheXpert	Stanford Hospital	English	224,316	0
CX-CHR	private dataset	Chinese	33,236	33,236
Padchest	Hospital Universitario de San Juan	Spanish	160,868	206,222
MIMIC-CXR	Beth Israel Deacones Medical Center	English	473,057	206,563

Table 1: Detailed information of different datasets

accuracy and robustness by using the SAE learning depth feature. (Avendi et al, 2016) (1)

Registration
The medical image registration task is also a common image analysis task. Basically, it tries to compare one medical image to another. For example, convert 3D images into 2D images. There is a research group using a 3D model developed to evaluate the pose of the 2D X-ray interpreter (Miao et al, 2016) (43). An image deformations prediction model was proposed by Yang et al (2016) (71), which used patch-wise image appearance from brain MRIs.

Other
Beyond the four approaches mentioned above, there is one more approach which not only takes medical images as input, but also could consider the textual information (the "report"). Essentially, it is a sub-task of classification. Some researchers think textual content could be combined with image information together as model input, then it could increase the model accuracy. In fact, they found when the model tried to identify various of pathologies in Optical Coherence Tomography (OCT) images, the classification accuracy was increased after they added semantic information mined from textual content (Schlegl et al, 2015) (47). Furthermore, a recurrent and convolutional network model was proposed by Shin et al (2016) (49), which trained both image and corresponding radiology reports together to classify anatomy, severity, and diseases. Another approach was to generate radiology reports from given chest X-ray images, partly influenced by Karpathy and Fei-Fei's work (2015)(29). One research group's approach used a topic vector model, which generated a topic vector first instead of directly generating words, then based on the topic vector, generated sentences. The model used a common encoder-decoder framework, CNN, as the encoder, and LSTM as decoder. It also introduced a co-attention mechanism which could help the model focus on important parts of the image (Jing et al, 2017) (27). Another group also applied Reinforcement Learning. Li et al (2018) (34) proposed a Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) model. The model considers either using a template database to generate sentences or using the LSTM model to generate sentences based on a given reward – a Consensus-based Image Description Evaluation score (CIDEr score).

2.2.3 Common chest X-ray dataset

At present, there are not many chest X-ray datasets available online. As shown in Table 1:

IU X-ray dataset
Demner et al (2015)(71) introduced IU X-ray dataset. The amount of images and reports is the smallest among these dataset, it has 7,470 chest X-ray images 3,955 and corresponding chest radiology reports. But it is the first time an institution released chest X-ray dataset is publicly available online. It collected data from Indiana Network for Patient Care. Due to the sensitive nature of the data and patient privacy, the entirety of the data is completely anonymized. All the individual's names have been replaced by "XXXX". Each file contains a patient's information, which include two types of X-ray image: both lateral, and a frontal view, and one report(annotation). The report has two sections: IMPRESSIONS and FINDINGS. In the impressions section, a brief diagnosis was provided by a radiologist, based on the patient's clinical record, and indications for the imaging study. In the findings section, the radiologist gives a list of observations and findings about the patient's condition, organ situation, etc.

There are many studies in the IU X-ray dataset. It has been used to test which pretrained CNNs architectures are most suited to the classification task by using IU X-ray dataset (Baltruschat et al, 2019) (2). The resutls showed that the best pretrained model was ResNet-38. As mentioned above, segmentation is an important task in medical image processing. The IU X-ray dataset was used in this field as well. Xnet was proposed by Bullock et al (2019) (4), which is an an end-to-end network. It could segment original images in to pieces, such as soft tissue, open beam regions,

and bone. One research group also used the IU X-ray dataset for the medical report generation task. Attention-based Abnormal-Aware Fusion Network (A3FN) was proposed by Xie et al. (2019) (66). It used corresponding patient images to generate logical radiology reports. More importantly, it could better concentrate on abnormal images, although it sometimes lacked accuracy.

Chest-Xray8 and Chest-Xray14 dataset
Wang et al (2017) (58) introduced the Chest-Xray8 dataset. Compared to the IU X-ray dataset, it is a much larger dataset. Wang et al (2017) only kept the front view image, and removed the lateral view from the dataset. They only recorded up to 8 common diseases for each image instead of the whole relevant radiology report. These diseases (labels) were extracted from corresponding original reports, then used one-hot encoding to represent each chest X-ray image. Consequently, a single image could include just one label, or it could be assigned multiple labels (diseases). Later the authors released a larger dataset called Chest-Xray14, which extends the labeling by an additional 6 common diseases for a total of 14 labels.

Several works has been done in this dataset. A 3D deep Convolutional Neural Network (CNN) was applied to identify breast cancer by Zhou et al (2019) (77). In terms of the segmentation task, an unsupervised domain adaptation method was proposed by Dou et al(2019) (14). They generalized the Convolutional Network by Generative Adversarial Nets (GAN) for medical images. There is also impressive classification research using this dataset. In order to distinguish both hip fractures and pelvic fractures, a weakly supervised method was proposed by Wang et al (2019) (61). It first used localized fractures with weakly supervised ROI mining.

CheXpert dataset
Irvin et al introduced CheXpert (Irvin et al, 2019)(24) which is similar to Chest-Xray14. It also includes 14 labels, however some labels are different. In order to label each report automatically, they designed a tool called "CheXpert labeler", which is specially designed to extract labels from radiological reports automatically. The Stanford ML group have provided us with the official split dataset, training, testing and validation sets, which have helped researchers to easily compare their results with other methods' results.

It also have several research group study on this dataset. Mostly is classification task since it does not have relevant reports. For example, a deep Hierarchical Multi-label Classification (HMLC) was proposed by Wei et al (2019) (62). It is not like the other systems mentioned, in that its models were trained with conditional probabilities, and then trained with unconditional probabilities. One research group used the CheXpert dataset as pretrained model for the classification task, and then also used multi-view image concepts to create a combined model to generate logical reports (Yuan et al, 2019). (74)

PadChest dataset
Bustos et al (2019) introduced PadChest dataset (5). It is another large chest X-ray dataset, which includes corresponding reports. There is not much research work that makes use of PadChest, one major reason being that PadChest is in Spanish, and is not an English language dataset. If PadChest were to be made available in English, we would likely see much greater use of it, as it is much larger than the previously mentioned datasets, with 160,868 images and 206,222 accompanying reports. It might be helpful if in the future researchers could make their datasets available in multiple languages.

CX-CHR dataset
Another dataset worth mentioning is the CX-CHR (34) dataset which is in Chinese. It has also not seen much use (there is currently only one research group making use of this data, which is the same group which collected and processed the data). The dataset includes 33,236 images and 33,236 corresponding reports selected from 35,500 patients, including multiple X-ray images of different views, such as posterior, anterior, and lateral views. The reports were preprocessed through tokenizing using a tool called Jieba, and filtered out vocabulary occurring under 3 times. The main issue with this dataset other than only being available in Chinese, is that it is a proprietary internal private dataset made by the same people who use the data for their research. Unless this dataset becomes open source, other groups will be unlikely to make use of this data.

MIMIC-CXR dataset
Johnson et al (2019) introduce MIMIC-CXR dataset. (5) The largest and the newest chest X-ray dataset publicly available online is MIMIC-CXR dataset. It uses the same labelling tool as CheXpert, and contains corresponding radiological

reports. Besides, it separate training testing validation dataset provided by the Beth Israel Deacones Medical Center. Researchers can easily compare their performance to the testing and validation dataset.

So far, only few works have been done using this dataset. Such as, tCheXnet, a model specifically designed for classifying pneumothorax based on chest X-ray images. Because it only focuses on one particular disease, the accuracy of this disease detection outperformed other models (Sze-To, A., Wang, Z, 2019) (53). It might lead to future research, using multiple classification models combining the results instead of only using one model to classify all diseases. One problem in medical report generation is sometimes the model will generate a ”plausible” report but not an ”accurate” one. Thus, an accuracy score should be considered when it generates a report, also combined with a readability score, the model achieved an impressive result in the end (Liu et al, 2019) (39). Another problem in the classification field is in abnormality classification. In real world practice, false positives would not be acceptable in medical report generation. Thus, how to choose a most suitable threshold for the model becomes a vital problem. (Wong, K. C., Moradi, M., Wu, J., Syeda-Mahmood, T. ,2019) (64)

2.2.4 Language Generation

Gatt, A., Krahmer, E., (2018) defined Language Generation tasks are text-to-text generation and data-to-text generation tasks. Both could be considered LG tasks.(18). There are several sub-fields, for example, image captioning, machine translation, question answering, summarization, etc. But the goal of this sub-task remains the same, which is the final generated textual content should be as real as possible, contains meaningful information, and last but not least it should to be linguistically accurate.

After Deep Learning was applied in this field, many impressing results were achieved. However, the development of Language Generation has been hindered for long time, the reason is that RNN architecture could decode a long meaningful sentence. Until Long Short-term Memory (LSTM) was proposed by Graves (2013) and achieved great results on the Wikipedia dataset (19), language generation tasks had not developed much. Since LSTM invented, it inspired all the relevant field. In terms of machine translation, traditionally the model applied RNN-RNN architecture, but some researchers approached a new architecture which used RNN as the encoder, LSTM as decoder(Sutskever et al, 2014) (51). This approach showed much better results than others. However, LSTM also has it own limitations. Some researchers found LSTM failed to generate documents that sufficiently resembled the sort of document a person would write, when new larger unseen datasets were introduced to the existing model. (Wiseman et al, 2017) (63)

2.2.5 Image Captioning

We also want to highlight some other topics relevant to possible future research. Image captioning, one type of Language generation model that not only takes text as input, but also consider images as input, is worth exploring further. Lin et al (2014) (37) introduced the dataset Microsoft COCO (MSCOCO). The dataset Microsoft COCO (MSCOCO) which is the largest natural image dataset. (Lin et al, 2014) (37) Each of its images is accompanied by five related annotations. Another two dataset worth mentioning are Flickr8k and Flickr30k, both are natural images datasets but much smaller than COCO dataset. In order to compare results effectively, there are four automatic evaluation methods have been used for image captioning, which include three traditional language generation’s automatic metrics such as BLEU, METEOR, ROUGE; and CIDEr which was especially designed for image captioning. This dataset attempts to improve object recognition by placing the object recognition question in context of broader information. The approach to achieving this was to gather images of complex every-day scenes that contained common objects in their common natural context. These images were labeled using per-instance segmentations to assist object localization precision. The dataset contains 2.5 million labeled instances in 328,000 images.

The first impressive result used CNN as encoder and RNN as decoder, as mentioned above, this kind of structure is same as the common structure used in machine translation. The different with machine translation is that the model takes image features as input instead of sentence features (Vinyals et al,2015)(55).

Because an image is basically a series of pixel values, the visual information could be converted into a matrix of values and then the corresponding visual features could be extracted from this data using a CNN, and then the features decoded into an output sequence. Essentially they took images as input, then attempted to generate readable, accurate sentences which could describe the images succinctly. However the sentences were overly simple in their descriptions, and did not sufficiently describe enough of the visual information.

Furthermore, this architecture has developed by Xu et al's group (2015) (67). They introduced an attention mechanism. Compared with Vinyals et al model, they devoted more attention to the Decoding stage with an attention mechanism. Basically, due to the attention mechanism, the decoder is able to select critical areas from the input image, which allows their model to generate a more detailed sentence during decoding, and has more words describing important features. For instance, as shown in Figure 1, it would describe more of the dogs instead of the grass.

These two articles above, both using last fully connected layer of CNN as "image semantics", but some researchers believed there are some high-level semantic features held in the CNN final classification layer. For example, "There is no dog", and other similar high-level semantic features could be greatly affected in the final generated result, and should not be discarded (Wu et al, 2016) (65).

Some researchers tried to study in the opposite direction. Their model not only could take the given image features and then decode them to be text data, but also to bi-directionally consider the text as it is being generated in the context of the image in order to dynamically compose text. In other words, they attempted to make the generated text as grounded in the visual features as possible in order to improve the quality of the output. Since the text generation was bi-directional, not only could an image be used to generate text, but text could be used to generate an image in order to reconstruct visual features. (Chen, X., Lawrence Zitnick, C., 2015) (8)

An Adaptive Attention structure was proposed by Lu, J., Xiong, C., Parikh, D., Socher, R. (2017). They took futher steps in developing the attention mechanism. All the models mentioned above have a common limitation, their decoder needs to generate word by word, even some preposition and function words. For those function words, prepositions and other minor parts of language can be inferred from the initially generated text. For example in the sentence "a yellow dog is running on the ground", if "dog" and "ground" could be predicted, then "a", "the", and "is" could then also subsequently be predicted. Their adaptive attention model could used both image and text information from existing captions, at each iteration of wording the caption, it considers whether to use the image information or not. If it used image information, it continues to use the attention model to generate word. Otherwise, the model will look at the existing generated text to predict next word. So, their model could use image information and text information alternatively.

2.2.6 Unsupervised Image Captioning

So far, this review covered most kinds of image captioning situations. All the models mentioned above required a paired image and text data. As discussed before, keeping a paired text and image data is quite expensive. In order to solve this problem, some researchers tried to explore unsupervised image captioning. Unsupervised image captioning was first proposed by Feng, Y., Ma, L., Liu, W., Luo, J. (2019).

This way, text datasets and image datasets can potentially come from different sources. They found visual concepts could be the link of image and text. And the visual concepts in their case is the objects in the image. Thus, their model needed to recognize all objects in the image, an object detection model is required. They first used a common object detection model to extract the objects from image dataset, and then they used these "visual concepts", fed into an adversarial model to determine in the generate sentence whether it has "visual concepts" or not. Their model used the Generate Adversarial Network (GAN) to generate a sentence, could distinguish whether the sentence belong to the corpus or not, and could also discriminate between whether the sentence cointained visual content or not. These two rewards – whether or not it is from the corpus, and whether not not it has visual content. They also designed a reconstruction model which could reconstruct the image from the generated sentence. However this approach did not show much improvement, with only about a one percent improvement when compared to models that do not use reconstruction.

2.2.7 Medical Report Generation

Deep learning for healthcare has been widely recognized as an area of high impact and has potential for both academia and industry. Automatic generation of medical imagery reports is gaining increasing research interest as one of the key applications in this field (27)(34)(32)(39). As shown in figure 1, if we compare the relatively simple and sufficient sentence 'a dog is running' to the images of dogs, and we look at the X-ray image and its accompanying text, we can see that the the essential facts of an X-ray image are a lot more nuanced and detailed, with a lot more information to be extracted from very subtle visual markers. Besides, even an experienced physician can misdiagnose or fail to completely identify all of the essential information. In other words, the raw medical report corpus is lacking an objective way to adequately judge and write descriptions, which makes it hard to generate an objective report based on just the raw corpus.

So far, a fair amount of research has been conducted in MRG field.

TieNet is a successful model which approached by Wang et al (2018) (59). It divided the image captioning task into several small tasks which include: multi-label classification, object detection, and report generation. The advantage is that it reduces the complexity of the whole image captioning model. As long as the model is sequentially processed, the current sub-module would take last sub-module results as input, and given results to the next-module. In the end, it could generate report successfully. Firstly, a given CT image would be used as input, then a pretrained CNN model used to extract image features. Afterwards, in order to get the feature maps, it has to go through an additional Transition Layer. Secondly, they used an attention mechanism to get text embedding by LSTM. Thirdly, Saliency Weighted Global Average Pooling was applied to get image embedding with textual information. Finally, their model could use image embedding and text embedding to predict disease labels. Their conclusion suggests that considering both kinds of information (image and text) could achieve the best results. Interestingly, the result indicated the best result was only 1 percentage better than using textual information alone.

Another research group also tried to divide the whole model into several sub-tasks. Their model more concentrated on the prediction module. They extracted keywords from medical reports, such as disease, organ, etc, which will be used as labels. Their first module is prediction or classification module, after parsing input chest X-ray images into a pretrained CNN, (in this case, they used ResNet) to get image features, they fed these features into a multi-label classification network (MLC) to predict labels they summarized before. Their second module is a generation module, their model would select the top A labels which have the highest probability scores. Then based on Tag Vocabulary, to get A labels' word embedding vector. Afterwards, a LSTM with attention mechanism will be used to get B context vectors and image features. Thus, B context vectors would generat B topic vectors, and B topic vectors would generate B sentences, each sentence corresponding to each topic vector. Additionaly, there is a stop control variable which could stop iteration when given conditions are satisfied (Jing, B., Xie, P., Xing, E., 2017) (27).

In the other direction, several research groups tried considering image captioning as a single process. For example, one group's approach used a hybrid model which partly relied on retrieval methods, and partly on generation methods. The reason is that for some rare diseases, the model performed better if using a retrieval method. In order to achieve this goal, they used reinforcement learning to select whether to use the template database or normal generative methods (Li et al, 2018) (34).

One research group tried to imitate the practical process a human practicioner goes through when writing a radiology report. The paper details how, according to their understanding, the radiologist first observes a chest X-ray image and obtains a brief descriptive sentence, for instance, "No acute disease". Then radiologist observes the heart area of chest X-ray image, and depending on what general observations were made in the first sentence, adds appropriate subsequent details in the second sentence (such as "the heart size is normal"). The lungs are then observed, in context of what was found earlier in the first and second sentences, and a third sentence is formulated (for example, "The lungs are clear"). This is how they broke down how a diagnostic report is written by a doctor. In reality, the reasoning of the doctor when writing a report may not resemble this observation at all and may follow faulty assumptions, but this is what the process appears to looik like to a non-medically trained observer, and at the very least gives us a sequential manner of thinking about the process. Whether this process matches what is happening in the doctor's mind may not be important if it gives us a valid pattern to follow. (Xue et al, 2018) (69)

Li et al (2019) (32) introduced a knowledge graph-driven model, the logic remaining the same as in their previous work. Their approach used a knowledge graph to establish connections among all the abnormalities. Their model contains two modules, a knowledge graph and retrieval module. It could present a sentence the from existing template database or use graph transfer to generate a new sentence.

Liu et al (39) took their own work further by attempting to solve how to increase the generated report accuracy in the medical image report generation field.

All the research mentioned evaluated their own performance using a readability score. It might guide the model to be biased towards generating sentences which superficially appear natural but are irrelevant in reality. Therefore, they concluded that classification accuracy still needed to be confirmed by comparing the generated report to the original Ground Truth (GT) report (using the same methodology used to compare labels). Because they considered both the classification accuracy and readability as important in their task, their work did not out perform the work of others.

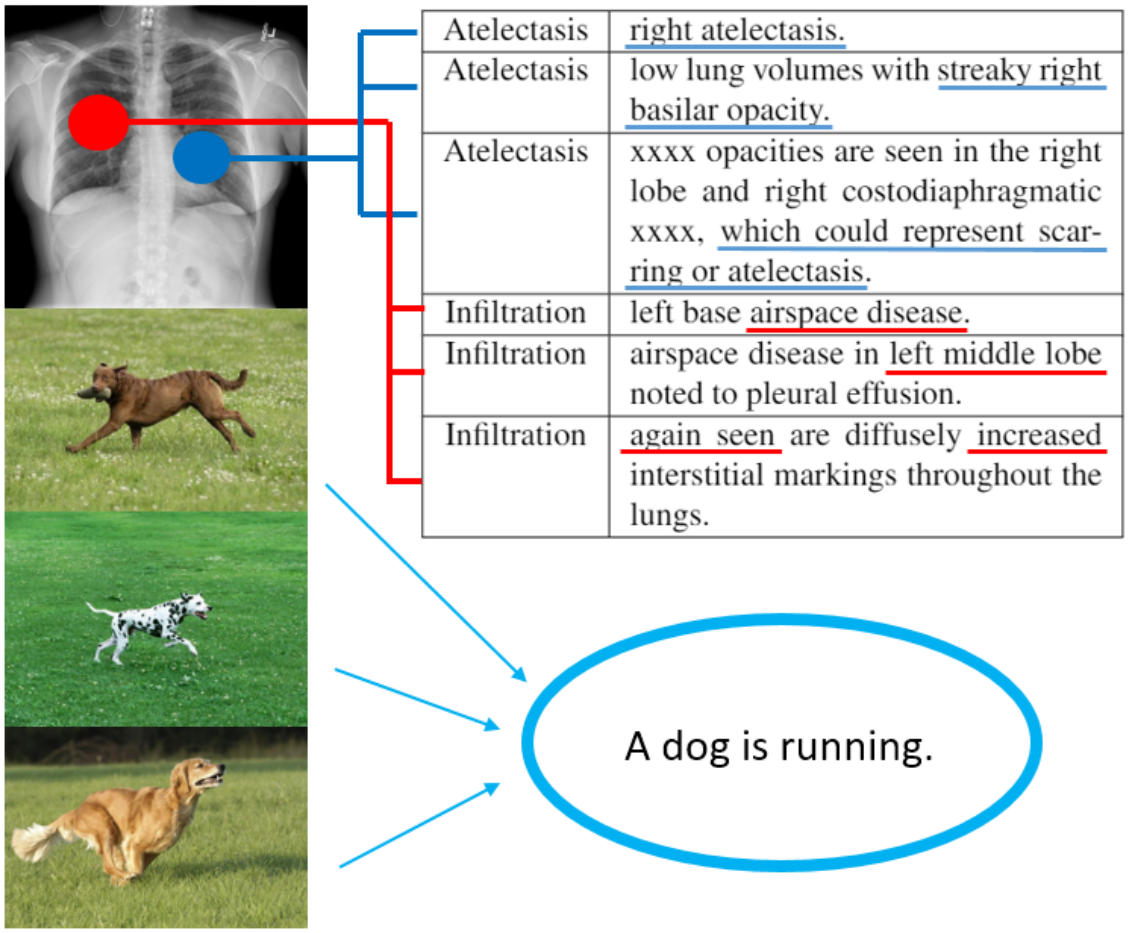

Figure 1: Compared to the task of traditional image captioning, medical report generation is more complicated as there is no typical ground truth for a single X-ray image. Description style ranges under different physicians. Some tend to be quite brief while others are more specific with logical reasoning of relative medical knowledge as well as including the patient's relevant medical history. This presents a big problem for fully supervised methods, as most image captioning models that have been adopted rely on strictly paired image-sentence data.

2.3 Summary

This review introduced and summarized the relevant fields of Medical Report Generation. From the development of Deep Learning in Radiology, to Language Generation, Image Captioning and Medical Report Generation. Based on the discussion above, firstly, the recent research mainly applied traditional supervised image captioning methods which required a paired image and report, more or less they have some bias toward current datasets. Secondly, MIMIC-CXR is a large, newly released dataset, only one research group has worked on this dataset, it contains more 'abnormal' images (images containing markers for conditions) which could help model the recognition of conditions. This research will try to apply a weakly supervised method on both IU X-ray and MIMIC-CXR dataset.

3 Chapter III

3.1 Introduction

The traditional image captioning approach which produces single-sentence descriptions has made impressive progress over the past few years(33)(42)(9)(76)(25)(26)(55)(72). Recently, long and semantic-coherent reports or stories have attracted more research interest with the more challenging but significant goal of bridging visual patterns and linguistic descriptions(35)(56)(41). As an emerging task of image based long text generation and radiology practice, the task of

medical image report generation is tougher under fully supervised methods(27)(34)(32)(39) for several reasons:

- Different physicians have different writing styles. As shown in Figure 1, there might be thousands of descriptive variations on one typical disease. Moreover, a patient's past medical history could influence the writing style a lot. Unfortunately this kind of information cannot be deduced from a single X-ray image.
- Noise is high in both images and reports. For images like chest X-ray images, the noise caused by different body shapes and standing postures presents a big problem for visual feature extraction from a greyscale image lacking 'clear' information. For reports, there are many invalid marks: For example 'xxxx' for sensitive information, or informal 'internal' abbreviations and incomplete sentences which can be deduced by human practitioners.
- Manually labeling medical images raises the already high work load for medical practitioners, causing the annotation to become inefficient and costly.
- There are various problems in the existing datasets. Using the IU X-ray dataset(12) as a benchmark for MRG(medical report generation) is too small to accurately extract and classify diseases through a CNN or a fine tuned pretrained CNN (e.g. ResNet(21), VGG(50)). Chest-Xray 14 dataset(57), based on which the CheXnet(45), a pretrained network providing predictions of 14 diseases, only contains paired image-disease labels with no report set. The same thing occured with another big dataset cheXpert(24), with 224,316 images without paired reports. Some datasets are in Spanish or Chinese(e.g. Padchest(5)), making the language barrier between datasets the reason for some of the limitations of MRG performance. Detailed information about these datasets can be found in Table 1. It is worth mentioning that the MIMIC-CXR dataset has still not been fully released, and so we have only used the report section in this thesis.

In this section, we make the first attempt to address a weakly supervised method of performing the MRG task. There are three components in our approach.

Firstly, the raw medical report corpus needs to be further subjected to selecting and cleaning. As illustrated in figure 2, the corpus filter model could remove the noise of reports (such as reports which contain a rare disease) greatly by using template embedding and clustering. Experiment shows that the model will accelerate the convergence of the sentence generation model and improve the accuracy of the final report according to different language metrics. At the same time, new concepts can be extracted from each sub-level to form a hierarchical concept pool. To ensure the final generated report was as readable and objective as we could manage, we kept only the most essential details from the corpus.

Secondly, inspired by MaskGan(17), we used an adversarial text generation method to train the language model on the corpus. As we do not set paired radiology report and chest X-ray images, we generated sentences using adversarial training. Thus, these sentences are indistinguishable from those in our medical report corpus.

Thirdly, the generated sentences need to describe the relative abnormalities and target diseases, we made a disease pool that included 14 common chest diseases. We also built sub-level hierarchical concepts for each disease. Semantic features from the pretrained Chexnet(45) were used as a top level for a hierarchical concept pool. An appropriate reward is granted according to the varying disease concept levels included in the sentences.

In summary, our contribution could be divided into four parts:

- We have made the first attempt to address medical report generation in the field of chest radiology using a weakly supervised approach.
- When compared to more traditional supervised approaches, we have displayed state-of-the-art report generation results, according to a variety of benchmark methods.
- Our proposed model is an effective approach that uses a hierarchical concept pool with noise reduction on large medical knowledge oriented corpus which improved the robustness of the final generated report.
- Our model uses a hierarchical concept pool, a method which has the potential for use in other image captioning domains characterised by the use of annotation and high cost, in particular, relating to complex fields of human knowledge.

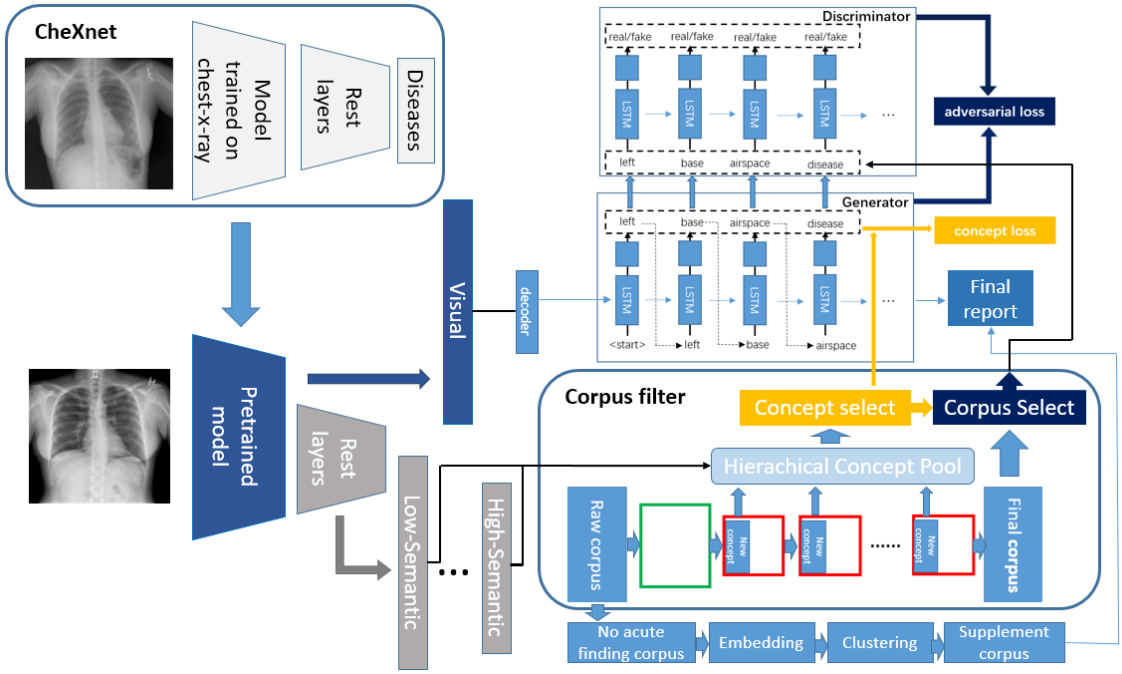

Figure 2: Illustration of our weakly supervised medical report generation model. Semantic features of 14 raw common diseases (high-semantic) with sub-level concepts (low-semantic) are used as select signals for a hierarchical concept pool. Visual features are adopted instead of high-level semantic features as input of GAN after FC decoder to preserve more information. Hierarchical concepts are extracted from the MIMIC-CXR raw corpus as the reward of sentences generated by GAN. The final report is the combination of a 'no acute finding' template and generated sentences.

3.2 Weakly Supervised Medical Report Generation

In this section, we will first discuss our corpus filter model, and we will then describe our novel weakly supervised medical report generating model based on a given chest X-ray image.

3.2.1 Corpus Filter Model

We first collected the reports from the MIMIC-CXR dataset and split the raw medical reports by comma. Thus we end up with a raw corpus of medical sentences called templates. These templates contain two distinct categories – positive diagnosis, or not, which we used to divide the corpus into two parts. The first part included templates with no acute disease findings (no positive findings). The second group contained templates with disease descriptors (positive diagnoses). We used the open-source auto labeler 'chexpert-labeler' to place items in one or the other group, and for disease detection to label the template as such. We used the normal template to generate a no acute finding report and mainly focused on disease description generation.

Next, we adopted an unsupervised method of reducing noise in the raw abnormal (positive diagnosis) templates. To ensure the final model generated objective and comprehensible results, we provided a template embedding method to eliminate scarce templates and clustering on the most used medical templates. First, we separated the raw abnormal template corpus into 14 groups according to prior work on chest X-ray image disease findings and (45)(20) scored each template using the remaining templates of the same group using a natural language metric commonly used in machine translation – Bleu1 – and excluded relatively meaningless words such as 'and' and 'is'. Second, we indexed all the templates in the same group, and for each template, we gave an n-dimension vector representing the template features where the value of k-dimension is the metric score between this template and the index-k template. In other words, the template feature embedding tool is to map the similarity between different templates on each dimension treating every template as unique. Then, we adopted the k-means++ approach to do clustering on the template group using template

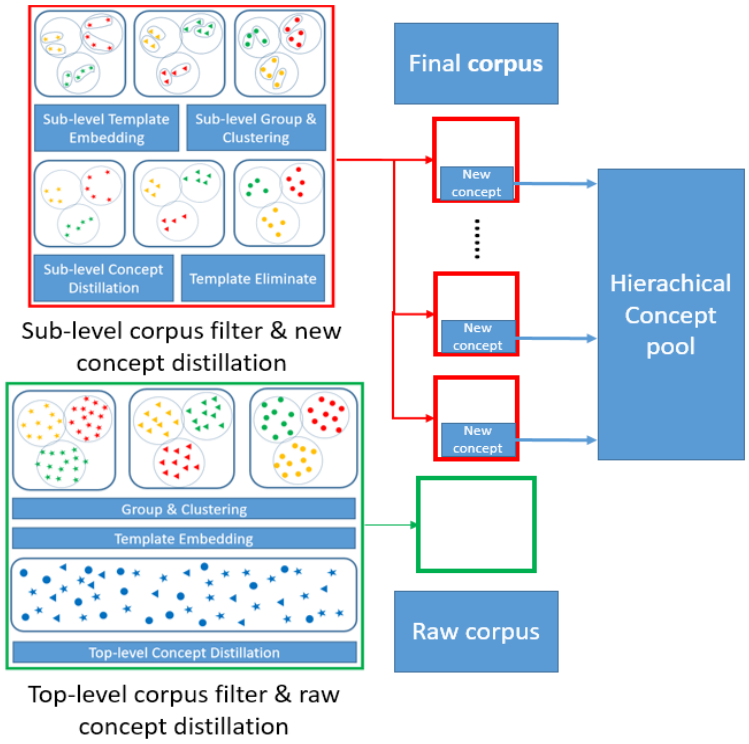

Figure 3: Illustration of corpus filter model for abnormalities. The final corpus will be used in medical report generation.

features. In this way we get sub-groups. For each sub-group, we computed the final score for each template by:

$$score^{k}_{final} = count_k + \lambda \cdot \sum_{i=0}^{n} (count_i \cdot score^{k \to i}_{similar}) \tag{1}$$

where n equals the size of the sub-group.

Finally, sorted by final score, we eliminated the last 70% of the templates in each sub-group to get the templates which are are commonly used. For the remaining templates, we counted word frequency, excluded meaningless words to distill sub-level concepts, and add the newly extracted concepts to the hierarchical concept pool. This step could be repeated several times on different tasks for different purposes. For medical report generation, if we prefer the final report to contain more medical reasoning or more specific descriptions, we would repeat the sub-level corpus filter several times to increase the depth of the hierchical concept pool with a higher reward placed on it. For this paper we only repeated this step twice to form a top-middle-bottom concept pool according to the raw corpus size and top-level concept number. The experiment results validate our assumptions. An instance of a hierarchical concept pool can be found in figure 4.

The last part of the corpus filter model is the noise reduction procedure for normal templates presenting no acute findings in a report, as the normal templates are all negative on disease findings. Thus, there is no disease group for the normal template corpus. For the readability of the generated report, we regard the whole normal template corpus as a group and performed clustering on it. Notice that for normal template clustering, the number of clustering nodes depends on the average template number each no acute finding report contained.

3.2.2 Medical Report Generation Model

As shown in Figure 2, the sentence generation model takes three components as input. The first input is the visual features of chest X-ray images $\mathbb{I}$, $I = I_1, I_2, ...I_{n_i}$, where n_i is the number of total images. Second is the template corpus input $\mathbb{S}$, $S = S_1, S_2, ...S_{n_s}$, where n_s is the size of the final corpus preprocessed by the corpus filter model. The third input is the semantic features provided by the pretrained CheXnet.

	Sub Group One	Sub Group Two	Sub Group Three
The number of templates	1541	1704	373
The number of selected templates	462	511	111
The max score template	Bibasilar atelectasis is mild	Minimal atelectasis at the left lung base	Atelectasis change are again seen at the left base
Example of selected template	Bibasilar atelectasis is also unchanged. Mild bibasilar atelectasis is also noted. Bibasilar atelectasis has improved.	Minimal atelectasis in the left lung base Minimal left basilar atelectasis is seen There is minimal atelectasis in the left lung base	Atelectasis change are also seen at the left base. Some Atelectasis changes are seen at the left base. Mild atelectasis changes are again seen at the left base.
Common Words	Bibasilar(437), Lung(305), Base(232), Mild(185), Minimal(138), Basilar(97), Noted(85), Unchanged(75)		
Middle level Concept	Bibasilar(437), Lung(305), Base(232)		

Figure 4: An instance of the middle layer of hierarchical concept pool at top level concept 'Atelectasis'. After getting high scoring templates, we extract the commonly used words by frequency from each sub-group and choose the most representative words 'Bibasilar', 'Lung', and 'Base' as middle level concept of 'Atelectasis'.

Image Encoder. Widely used pretrained Convolutional Neural Networks (e.g. ResNet(21), VGG(50)) perform poorly on X-ray images for the large domain gap between a 'lovely cat' and a 'chest X-ray'. To overcome this, CheXnet (45)is proposed. Trained on ChestX-ray14 which contains 108,948 chest X-ray images with paired diseases, CheXnet can provide disease prediction with acceptable accuracy. Thus, the features extracted from chest X-ray images could be represented by F_{cxim}

$$F_{cxim} = CheXnet(I) \tag{2}$$

Here we fine tune the CheXnet with two hierarchical concept levels followed by each top-level disease concept.

Sentence Generator. We use Long short-term memory (LSTM) as the generator, it takes F_cxim and previously generated words as input to generate sentences word-by-word until it reaches the end-of-token. As shown in Figure 2, at each time-step the output of LSTM is the word that has the highest conditional probability at the current state.

$$\mathbf{x}_{-1} = FCL(F_{cxim}) \tag{3}$$

$$\mathbf{x}_t = \mathbf{W}_e \mathbf{s}_t, t \in \{0...n-1\} \tag{4}$$

$$\left[\mathbf{p}_{t+1}, \mathbf{h}^g_{t+1}\right] = \mathbf{LSTM}^g(\mathbf{x}_t, \mathbf{h}^g_t), t \in \{-1...n-1\} \tag{5}$$

$$\mathbf{s}_t \sim \mathbf{p}_t, t \in \{1...n\} \tag{6}$$

FCL represents fully-connected layer, We feed the image feature into a 512 dimension fully connected layer to get $x_($ $-$ $1)$, W_e is the word embedding matrix, at t-th time step s_t is one-hot vector of generated words. Multiply both and we get x_t which is the input of LSTM except at $t = -1$ $s0$ and sn denote the start-of-sentence (SOS) and end-of-sentence (EOS) tokens, respectively. h^g_t is the hidden state of LSTM, we set it as zero when $t = -1$.The output of LSTM at each step is p_{t+1}, it represents the probability over the dictionary at the t-th time step.

Sentence Discriminator. We also used LSTM as the discriminator. It tries to distinguish whether the sentence is from the template or the new sentence was generated by the model itself.

$$[\mathbf{q}_t, \mathbf{h}^g_t] = \mathbf{LSTM}^d(\mathbf{x}_t, \mathbf{h}^d_{t-1}), t \in \{1...n\} \tag{7}$$

$\mathbf{q}_t$ represents the probability of the generated sentence $\mathbf{s}_t = [\mathbf{s}_1...\mathbf{s}_t]$ is regarded as real by the discriminator. $\mathbf{h}^d_t$ denotes LSTM hidden state.

Adversarial Sentence Generation In order to train our weakly supervised medical report generation model, since we did not have paired image-sentence data available, we attempted to generate pseudo sentences, and used them to train in a supervised learning manner. Our approach employed two methods to make sure the pseudo sentences were as real as possible. As discussed before, our model takes chest X-ray image features and semantic features as input to generate sentences word by word. Since there is no ground truth sentence, the word is generated on the current probability distribution which could lead the model to generate nonsensical sentences. In order to fix this problem, an adversarial reward method was proposed. After generating a sentence, the discriminator decides whether the sentence is from the existing corpus (real) or if it is a new sentence generated by the model itself (fake). In this way, the generator was forced

to generate sentences that appear as real as possible. Next, we gave a reward to the generator at each time-step. The reward value for the t-th generated word is the logarithm of the probability estimated by the discriminator:

$$r_t^{adv} = log(q_t) \tag{8}$$

During the training process, the generator would learn to generate plausible sentences by maximizing the adversarial reward. For the discriminator, the corresponding adversarial loss is defined as:

$$\mathbf{L}_{adv} = -\left[\frac{1}{l}\sum_{t=1}^{l} log(q_t) + \frac{1}{n}\sum_{t=1}^{n} log(1-q_t)\right] \tag{9}$$

Hierarchical Concept Reward The problem of using the adversarial reward is that it only guarantees that the generated sentence is plausible, but the sentence might be irrelevant to the image itself. In order to fix this problem, we proposed a hierarchical concept reward model to constrain the generator. The 14 most common diseases such as 'Atelectasis', 'Cardiomegaly' and 'Effusion' form the top-level concepts. Following each top-level concept, we add 2 to 5 of the most common concepts extracted from each sub-group using the corpus filter model, such as 'lobe', 'lung', and 'pleural' as middle-level concepts. Finally, we added an extra 5-10 concepts, such as 'area', 'linear', and 'trace' as the bottom-level concepts in the same manner. If the model generates a word which corresponds to these concepts, we give a reward to the generated word with the value indicated by the confidence score of that concept. In this way, we can generate more accurate sentences which are not only plausible but also relevant to the given image.

For a chest X-ray image I_i, the semantic features are the set of visual concepts given by CheXnet: $C = (c_1, cs_1), ..., (c_i, cs_i), ..., (c_{N_c}, cs$ where c_i is the i-th detected visual concept, cs_i is the confidence score associated with the related concept, and N_c is the total number of visual concepts. The concept reward assigned to the t-th generated word s_t is given by:

$$r_t^c = \sum_{i=1}^{N_c} \mathbf{I}(s_t = c_t) * sc_i \tag{10}$$

where $\mathbf{I}()$ is the indicator function.

Integration These two methods were considered when we trained the sentence generation model. For the generator, we trained the generator using the policy gradient method (52), because the word sampling operation was not differentiable. Policy gradient estimates the gradients with respect to trainable parameters given the adversarial reward and concept reward. Besides the gradients estimated by the policy gradient. We set θ to denote the trainable parameters in the generator. The gradient with respect to θ is given by:

$$\nabla_\theta \mathbf{L}(\theta) = -\mathbb{E} - \left[\sum_{t=1}^{n}(\sum_{s=t}^{n} \gamma^s (r_s^{adv} + \lambda_c r_s^c) - b_t)\nabla_\theta log(\mathbf{s}_t \mathbf{p}_t)\right] \tag{11}$$

where γ is a decay factor, and b_t is the baseline reward estimated using self-critique (46). λ_c are the hyperparameters controlling the weights of different terms. For the discriminator, the adversarial loss is used to update the parameters via gradient descent:

$$\mathbf{L}_D = \mathbf{L}_{adv} \tag{12}$$

During our training process, the generator and discriminator are updated alternatively.

3.3 Experiment

3.3.1 Dataset

In our model we collect images and radiology reports from two different public available chest X-ray datasets.

IU X-Ray We used the IU X-Ray (Indiana University Chest X-Ray dataset) as the image set consisting of 3,996 chest X-ray images, including both frontal view and lateral view. In order to make the results more comparable, we split the IU X-ray dataset in the same manner as (32), with 5,239 X-ray images as the training set, and 754 X-ray images for validation. The remaining X-ray images were used for testing. After data cleaning, we kept 2,217 chest X-ray images in

	Method(Paired)	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	ROUGE	CIDEr
Supervised	1-NN	0.232	0.116	0.051	0.018	0.201	0.113
	S,T	0.265	0.157	0.105	0.071	0.306	0.110
	S,A,T	0.313	0.198	0.135	0.103	0.273	0.155
	TieNet	0.330	0.194	0.124	0.081	0.291	0.184
	HRGR-Agent	**0.438**	**0.298**	**0.208**	**0.151**	**0.322**	**0.343**
	Method(Unpaired)	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	ROUGE	CIDEr
Ours(Weakly supervised)	rc/top	0.332	0.253	0.142	0.118	0.230	0.183
	rc/middle	0.341	0.260	0.146	0.120	0.243	0.197
	rc/bottom	0.320	0.231	0.125	0.109	0.215	0.171
	nr/top	0.360	0.265	0.154	0.124	0.256	0.201
	nr/middle	**0.397**	**0.282**	**0.195**	**0.147**	**0.287**	**0.257**
	nr/bottom	0.334	0.236	0.144	0.096	0.234	0.183

Table 2: Baseline models of the IU X-ray dataset at the upper section with our weakly supervised approach of the IU X-ray dataset at lower section. BLEU-n counts up n-gram for evaluation. Our approach is better than most baseline models except for HRGR-Agent trained on both the IU X-ray and a private dataset, CX-CHR.

the training set, 636 in the testing set, and 323 in the validation set.

MIMIC-CXR We collected the medical report corpus from MIMIC-CXR, the largest radiology dataset available, consisting of 206,563 reports from 63,478 patients. The radiological reports were parsed into sections, from which we extracted the findings section. We used the open-source auto labeler 'chexpert-labeler' (24) to label the diseases on reports from MIMIC-CXR and split reports in to two groups: 52, 544 normal reports and 154, 019 abnormal reports, the later group consisting of at least one kind of disease.

3.3.2 Implementation details

We split each report into sentences by comma and arrived at 32,959 normal sentences and 16,317 abnormal sentences. The abnormal templates were divided into 14 groups, followed by top-level concepts. After template elimination was performed on the corpus filter model, we arrived at 1,318 normal templates and 5,223 abnormal templates, which we used to build the final corpus of the sentence generation model.

We tokenized all of the words from the corpus to build the vocabulary. There are 663 words in our vocabulary, including special tokens: SOS, EOS, and an Unknown token. The LSTM hidden states were fixed to 512 dimensions. The weighting hyperparameters were chosen to make different rewards roughly at the same scale. Specifically, λ_c is set to be 10, γ is set to be 0.9. We trained our model using the Adam optimizer (31), with a learning rate of 0.0001. When generating the sentence in the test phase, we used beam search with a beam size of 3.

We applied softmax on the prediction and took the top k as the final diseases. Here k depends on the positive prediction of diseases, we set k to 2 in most cases only if the number of positive predictions was larger than 8 out of 14, and we set k to 3. At the middle level, the confidence score was set to 0.4, and 0.2 at the bottom level.

3.3.3 Baselines

We first compared our method with a KNN model and set k=1, '1-NN' where we used the test image to query the closest neighbors in the training set, in the image embedding space. The corresponding report of the nearest neighbor is used as the output of this test image. We then compared our model with the two image captioning model 'S,T' (Show and Tell) (55) and S,A,T' (Show, Attend, and Tell) (68). We also compared with two medical report generation models: TieNet (60) and HRGR-Agent (34). Additionally, to show the effectiveness of the hierarchical concept reward mechanism and noise reduction techniques, we also implemented six ablated versions of our model:'rc/top', 'rc/middle', 'rc/bottom', 'nr/top', 'nr/middle', and 'nr/bottom'. 'rc' indicates the row corpus without noise reduction, and 'nr' represents noise reduction applied. 'top', 'middle', and 'bottom' are the levels of concepts used from the hierarchy concept pool.

		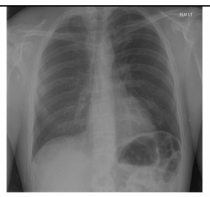		
labeler	No finding	Pneumothorax	Edema, Cardiomegaly, Atelectasis	Atelectasis
CheXnet	No finding	Edema, Cardiomegaly	Cardiomegaly, Atelectasis, Pneumonia	Effusion, Atelectasis
Ground Truth	No acute cardiopulmonary abnormality. The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation.	Cardiomediastinal silhouette is within normal limits. There is a minimally displaced right lateral th rib fracture and probable nondisplaced right lateral th rib fracture. There is a moderate right-sided pneumothorax measuring approximately . cm in the right apex. Moderate right-sided pneumothorax measuring approximately . Minimally displaced right lateral th rib fracture probable nondisplaced right lateral th rib fracture. Left lung is clear.	Stable subsegmental bibasilar atelectasis. Stable enlarged cardiomediastinal silhouette. Cardiomegaly without heart failure. Nasogastric tube tip xxxx within the stomach body, stable. Atherosclerosis of the thoracic aorta. Relative elevation of right hemidiaphragm. No focal consolidation, pneumothorax or large pleural effusion. Stable obscuration of lateral left diaphragm. Low lung volumes.	Low lung volumes on the ap view with bronchovascular crowding and bibasilar atelectasis. No focal airspace consolidation, pleural effusions or pneumothorax. No acute bony abnormalities. The heart size and mediastinal contours appear within normal limits. No acute cardiopulmonary findings.
TieNet	Pa and lateral views of the chest demonstrate lungs that are clear without focal mass , infiltrate or effusion . Cardiomediastinal silhouette is normal size and contour. Pulmonary vascularity is normal in caliber and distribution. No evidence of acute pulmonary pathology	Comparison is to previous upright study of no significant interval change is seen in the appearance of the chest. Cardiomegaly and congestive heart failure, with less severe interstitial edema. No evidence of acute pulmonary pathology	A single ap view of the chest demonstrates unchanged bilateral reticular opacities, consider atelectasis. No focal consolidation concerning for pneumonia. Heart size is mildly enlarged. no evidence of developing infiltrate .	Opacification at the right base is consistent with atelectasis. There are small bilateral pleural effusion. Endotracheal tube and ng tube have been removed. No pneumothorax. no evidence of acute pulmonary pathology
Ours(nr/top)	There is no pleural effusion or pneumothorax. The heart is normal in size. No pneumothorax. No acute cardiopulmonary abnormality. Cardiac and mediastinal contours are within normal limit. The lungs are clear	A component of edema is difficult to exclude. Cardiomegaly is mild to moderate. No focal infiltrates. There is no pneumothorax or pleural effusion. There are low lung volumes	However, in the appropriate clinical setting, superimposed pneumonia could be consider. Borderline cardiomegaly. Mild atelectasis changes are again seen at the left base. There is no pleural effusion or pneumothorax. The lungs are clear.	Bibasilar opacifications are consistent with small pleural effusion. Bibasilar atelectasis is mild. No pneumothorax. No acute cardiopulmonary abnormality. The heart size still appear within normal limits .
Our(nr/middle)	There is no pleural effusion or pneumothorax. The heart is normal in size. No pneumothorax. No acute cardiopulmonary abnormality. Cardiac and mediastinal contours are within normal limit. The lungs are clear	Cardiomegaly with early pulmonary edema, likely secondary to congestive heart failure. The heart size is borderline for cardiomegaly, cardiac enlargement. No focal infiltrates. There is no pneumothorax or pleural effusion. There are low lung volumes	New right lower lobe opacity concerning for pneumonia. Again seen is cardiomegaly, cardiac silhouette is borderline enlarged. There are relatively low lung volumes and bibasilar atelectasis. There is no pleural effusion or pneumothorax. The lungs are clear	Apart from mild bibasilar atelectasis, the lungs are clear without focal consolidation. Bilateral pleural effusion are noted, left greater than right. No pneumothorax. No acute cardiopulmonary abnormality. The heart size still appear within normal limits .

Figure 5: The qualitative results of our model 'nr/top''nr/middle' and TieNet. Text in the same color indicate the use of concepts of the same top level concept. We do not provide the generation of 'nr/bottom' and 'rc' due to too much redundant content.

3.4 Evaluation

Concept Prediction As our medical corpus contains no reports from the IU X-ray dataset, we assessed the top-level concept prediction accuracy on the IU X-ray image set using raw disease labels from the IU X-ray report. We got the AUROC of each disease and computed the F1 score of each concept prediction. The F1-score is the harmonic average of the precision and recall of the model. We used it as the metric to select a suitable threshold on each concept prediction. Here we reported the performance on both ChestX-ray14(pretrained) and our IU X-ray dataset in table 3.

Natural Language Metrics We reported the medical report generation results using the following captioning evaluation tools: $BLEU$ (44), $ROUGE$ (36), $CIDEr$ (54) which computed using the coco-caption code [1] For supervised medical report generation, as shown in the upper part of Table 2, it is clear that the HRGR-Agent (34) model had the best performance. At the lower part of Table 2, it can be observed from the results that the middle level had better performance than the others, 'nr/middle' is 5% higher than 'nr/top' at $CIDEr$ score , 'rc/middle' is 1.4% higher than 'nr/top' at $CIDEr$ score, the reason being that we could take more concepts as reward at the middle level than at the top level. Thus, it could generate a more specific sentence. At the bottom level, both evaluation scores using 'rc' and 'nr'

[1] https://github.com/tylin/coco-caption

	AUROC@mean	F1@max+mean.
ChestX-ray14	0.821	0.751
IU X-ray	0.792	0.725

Table 3: Prediction Accuracy on pretrained CheXnet. Despite the domain gap between two datasets, CheXnet still performed well on IU X-ray compared to ChestX-ray14.

decreased 2.6% and 7.4% at $CIDEr$ score due to excessive concepts which misled the model to generate redundant content. The best result lied in the middle level concepts with the noise reduction method applied on different metrics. Although our method performed better than some baseline models, it still falls behind the HRGR-Agent (34) model, however we should keep in mind that the HRGR-Agent was trained on a private dataset and implemented in a fully supervised manner.

Captioning Results Figure 5 demonstrates the qualitative results of our model 'nr/top', 'nr/middle', and TieNet. In general, 'nr/middle' had superior performancecompared to the other two methods. In terms of normal reports, all three models performed well, with several "no findings" sentences such as "no cardiopulmonary" and "lungs are clear". In case 2, since our model's performance partly relied on the concepts provided by CheXnet, there was deviation on the prediction, which lead to irrelevant sentence generation. In case 3 and 4, for the report with labels like 'atelectasis' or 'effusion', our model could generate reports with more relevant content when compared to TieNet reports. Here we notice that the multi-level concepts do help to improve report generation. Vinyals2015Show As shown in case 4, we have two extra concepts 'bibasilar' and 'lung' at the middle level. Our model was forced to generate sentences to get more concept rewards. At 'nr/middle', our model generated the sentence "Apart from mild bibasilar atelectasis, the lungs are clear without focal consolidation." While at 'nr/top', the model generated "Bibasilar atelectasis is mild". Compared with the ground truth "Low lung volumes on the ap view with bronchovascular crowding and bibasilar atelectasis.", our 'nr/middle' result was more accurate than the original sentence. Besides, we should note that most generated reports are shorter than the original human written reports. The reason lies in the absense of suggestion and inference which is hard to generate without knowledge of other medical domains.

Human Evaluation We randomly selected 10% of the data from the test set and invited a clinical expert to evaluate the generated reports of TieNet and our own. Several conclusions were drawn. First, our 'nr/middle' method could generate more detailed and correct information than TieNet. Second, reports generated using our method were more structured than TieNet, but less structured than the original ground truth. Third, all three models failed to generate an impression sentence.

4 Chapter IV

4.1 Conclusion and Future work

In this thesis, we presented a weakly supervised medical report generation model to automatically generate reports for medical images. Our proposed method addressed two major challenges: 1) how to overcome the limitations of released datasets lacking paired image-report data. 2) How to form a hierchical concept pool to guide the report generation. To deal with these two challenges, we introduced a noise reduction method with a concept distillation mechanism. We utilized these hierarchical concepts as signals to get rid of paired data and focused on disease detection and more practical description. On the widely used IU X-ray dataset, our proposed model was able to generate quite promising medical reports and show state of the art results, according to a variety of report benchmarks.

In the future, we plan to do more in the area of transfer learning, in order to assess our model's performance with other datasets. Our model had some limitations which need to be addressed in the future, such as, dependency on using a weakly supervised approach. We would like to be able to use a fully unsupervised approach in the future.

References

[1] MR Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical image analysis*, 30:108–119, 2016.

[2] Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019.
[3] Avi Ben-Cohen, Idit Diamant, Eyal Klang, Michal Amitai, and Hayit Greenspan. Fully convolutional network for liver segmentation and lesions detection. In *Deep learning and data labeling for medical applications*, pages 77–85. Springer, 2016.
[4] Joseph Bullock, Carolina Cuesta-Lázaro, and Arnau Quera-Bofarull. Xnet: A convolutional neural network (cnn) implementation for medical x-ray image segmentation suitable for small datasets. In *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10953, page 109531Z. International Society for Optics and Photonics, 2019.
[5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*, 2019.
[6] Gustavo Carneiro, Jacinto C Nascimento, and António Freitas. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. *IEEE Transactions on Image Processing*, 21(3):968–982, 2011.
[7] Gabriel Chartrand, Phillip M Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J Pal, Samuel Kadoury, and An Tang. Deep learning: a primer for radiologists. *Radiographics*, 37(7):2113–2131, 2017.
[8] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.
[9] Xinpeng Chen, Ma Lin, Wenhao Jiang, Yao Jian, and Liu Wei. Regularizing rnns for caption generation by reconstructing the past with the present. 2018.
[10] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.
[11] Bob D de Vos, Jelmer M Wolterink, Pim A de Jong, Max A Viergever, and Ivana Išgum. 2d image classification for 3d anatomy localization: employing deep convolutional neural networks. In *Medical imaging 2016: Image processing*, volume 9784, page 97841Y. International Society for Optics and Photonics, 2016.
[12] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.
[13] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
[14] Qi Dou, Cheng Chen, Cheng Ouyang, Hao Chen, and Pheng Ann Heng. Unsupervised domain adaptation of convnets for medical image segmentation via adversarial learning. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, pages 93–115. Springer, 2019.
[15] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent CT Mok, Lin Shi, and Pheng-Ann Heng. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.
[16] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
[17] William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: better text generation via filling in the_. *arXiv preprint arXiv:1801.07736*, 2018.
[18] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.
[19] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
[20] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, and Yang Yi. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. 2018.
[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
[23] Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–246. Springer, 2016.
[24] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.
[25] Wenhao Jiang, Ma Lin, Xinpeng Chen, Hanwang Zhang, and Liu Wei. Learning to guide decoding for image captioning. 2018.
[26] Wenhao Jiang, Lin Ma, Yu Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. 2018.
[27] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.
[28] Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856, 2009.
[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
[30] Jeremy Kawahara and Ghassan Hamarneh. Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In *International workshop on machine learning in medical imaging*, pages 164–171. Springer, 2016.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
[32] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673, 2019.
[33] Lijun Li and Boqing Gong. End-to-end video captioning with multitask reinforcement learning. 2018.
[34] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems*, pages 1530–1540, 2018.
[35] Xiaodan Liang, Zhiting Hu, Zhang Hao, Chuang Gan, and Eric P. Xing. Recurrent topic-transition gan for visual paragraph generation. 2017.
[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004.
[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
[38] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
[39] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*, 2019.
[40] Jiamin Liu, David Wang, Le Lu, Zhuoshi Wei, Lauren Kim, Evrim B Turkbey, Berkman Sahiner, Nicholas A Petrick, and Ronald M Summers. Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Medical physics*, 44(9):4630–4642, 2017.
[41] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
[42] Chen Long, Hanwang Zhang, Jun Xiao, Liqiang Nie, and Tat Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[43] Shun Miao, Z Jane Wang, and Rui Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE transactions on medical imaging*, 35(5):1352–1363, 2016.
[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL '02*, pages 311–318, Morristown, NJ, USA, 2001. ACL.
[45] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
[46] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. 2016.
[47] Thomas Schlegl, Sebastian M Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.
[48] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015.
[49] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.
[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[51] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
[52] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.
[53] Antonio Sze-To and Zihe Wang. tchexnet: Detecting pneumothorax on chest x-ray images using deep transfer learning. In *International Conference on Image Analysis and Recognition*, pages 325–332. Springer, 2019.
[54] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society, 2015.
[55] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
[56] Xin Wang, Wenhu Chen, Yuan Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. 2018.
[57] X Wang, Y Peng, L Lu, Z Lu, M Bagheri, and RM Summers. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, 2017.
[58] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
[59] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common

thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[60] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, pages 9049–9058. IEEE Computer Society, 2018.

[61] Yirui Wang, Le Lu, Chi-Tung Cheng, Dakai Jin, Adam P Harrison, Jing Xiao, Chien-Hung Liao, and Shun Miao. Weakly supervised universal fracture detection in pelvic x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 459–467. Springer, 2019.

[62] Jason W Wei, Arief A Suriawinata, Louis J Vaickus, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Naofumi Tomita, Behnaz Abdollahi, Adam S Kim, Dale C Snover, et al. Deep neural networks for automated classification of colorectal polyps on histopathology slides: A multi-institutional evaluation. *arXiv preprint arXiv:1909.12959*, 2019.

[63] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.

[64] Ken CL Wong, Mehdi Moradi, Joy Wu, and Tanveer Syeda-Mahmood. Identifying disease-free chest x-ray images with deep transfer learning. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 109500P. International Society for Optics and Photonics, 2019.

[65] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212, 2016.

[66] Xiancheng Xie, Yun Xiong, S Yu Philip, Kangan Li, Suhua Zhang, and Yangyong Zhu. Attention-based abnormal-aware fusion network for radiology report generation. In *International Conference on Database Systems for Advanced Applications*, pages 448–452. Springer, 2019.

[67] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[68] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015. cite arxiv:1502.03044.

[69] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466. Springer, 2018.

[70] Dong Yang, Shaoting Zhang, Zhennan Yan, Chaowei Tan, Kang Li, and Dimitris Metaxas. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 17–21. IEEE, 2015.

[71] Xiao Yang, Roland Kwitt, and Marc Niethammer. Fast predictive image registration. In *Deep Learning and Data Labeling for Medical Applications*, pages 48–57. Springer, 2016.

[72] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Mei Tao. Boosting image captioning with attributes. 2016.

[73] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4709–4717, 2017.

[74] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer, 2019.

[75] Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–328. Springer, 2017.

[76] Gan Zhe, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Deng Li. Semantic compositional networks for visual captioning. 2017.

[77] Juan Zhou, Lu-Yang Luo, Qi Dou, Hao Chen, Cheng Chen, Gong-Jie Li, Ze-Fei Jiang, and Pheng-Ann Heng. Weakly supervised 3d deep learning for breast cancer classification and localization of the lesions in mr images. *Journal of Magnetic Resonance Imaging*, 50(4):1144–1151, 2019.