Addressing the Challenge of limited availability of Properly Annotated and Categorised Training Data in Medical VQA Task

A

report submitted in partial fulfillment for the award of the degree of

Integrated Post Graduate Masters of Business Administration

in

Information Technology

By

Nirupama Singh : 2020IMG-045

Under the Supervision of

Dr. Vinay Singh

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
AND MANAGEMENT GWALIOR
GWALIOR, INDIA

DECLARATION

I hereby certify that the work, which is being presented in the report/thesis, entitled Addressing the Challenge of limited availability of Properly Annotated and Categorised Training Data in Medical VQA Task, in fulfillment of the requirement for the award of the degree of Bachelor of Technology and submitted to the institution is an authentic record of my/our own work carried out during the period May-2023 to August-2023 under the supervision of Dr. Vinay Singh. I also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Dated: **Signature of the candidate**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dated: **Signature of supervisor**

Acknowledgements

I am highly indebted to Dr. Vinay Singh , for his esteemed mentorship, and for allowing me to freely explore and experiment with various ideas in the course of making this project a reality. The leeway I was given went a long way towards helping cultivate a genuine hunger for knowledge and keeping up the motivation to achieve the best possible outcome. I can genuinely say that this Bachelor's Thesis Project (BTP) made me explore many areas of machine learning that are new to me, and kindled an interest to further follow up on some of those areas. Moreover, the semi-successful completion of this project has brought with it great satisfaction and more importantly, confidence in my ability to produce more high-quality non-trivial artificial intelligence systems that can make a difference in the real-world. I would like to sincerely express my gratitude to this prestigious institution for providing me and my colleagues with the opportunity to pursue this BTP. It is an honor to be able to work on such an important academic project under the guidance and support I am provided with. I am grateful for the resources and facilities provided by this institution, which have been instrumental in enabling me to conduct my research and complete this project. Moreover, I deeply appreciate the efforts of my professors in mentoring and fairly evaluating our works.

Nirupama Singh

Abstract

In the field of medical diagnostics and healthcare, there is a growing interest in combining artificial intelligence with visual information. One promising area is the Medical Visual Question Answering (VQA) task, where AI answers questions about medical images. This can improve patient care and diagnosis. However, the success of these systems depends on having good training data with accurate annotations. This Bachelor's Thesis Project addresses the challenge of not having enough high-quality training data for medical VQA.

The proposed approach combines two methods: the unsupervised Denoising Auto-Encoder (DAE) and supervised Meta-Learning. The Denoising Auto-Encoder (DAE) is used to handle the large amount of unlabeled images. It learns to create clear images from noisy ones. This helps the model learn important patterns in the data even without labels.

At the same time, the framework uses Meta-Learning to create adaptable weights for the model. This technique uses a small amount of labeled data to develop these adaptable weights. These weights help the model work well even when there isn't much labeled VQA data available.

Keywords: *Unsupervised Denoising Auto-Encoder, Meta-Learning, Medical VQA, Deep Learning*

Contents

List of Figures

List of Acronyms

CNN	Convolutional Neural Network
ML	Machine Learning
MSE	Mean Squared Error
MAML	Model-Agnostic meta-Learning
CDAE	Collaborative Denoising Autoencoder

1

Introduction

1.1 Introduction

Medical Visual Question Answering (VQA) is a cutting-edge interdisciplinary field that lies at the intersection of computer vision, natural language processing (NLP), and healthcare. It represents a promising area of research that leverages the power of artificial intelligence (AI) to bridge the gap between medical images and human-readable explanations. As the healthcare industry continuously embraces digitalization and technology ad- vancements, the integration of AI-based systems like VQA holds tremendous potential for enhancing medical diagnostics, decision-making, and patient care. This task is highly challenging due to the complex and heterogeneous nature of med- ical images, which often contain subtle abnormalities, anatomical variations, and intricate structures. Additionally, generating accurate responses to textual questions requires a deep understanding of both medical content and natural language semantics.

However, the success of VQA systems hinges on the availability of high-quality training data that accurately reflects the intricate relationship between medical images and associated queries. This requirement poses a significant challenge, especially in the context of medical VQA, where both visual and textual components must be understood and integrated effectively. The "Addressing the Challenge of Limited Availability of Properly Annotated and Categorized Training Data in Medical VQA Task" project delves into a critical aspect of this challenge, aiming to explore innovative strategies to mitigate the scarcity of well-annotated and properly categorized training data in the medical VQA landscape.

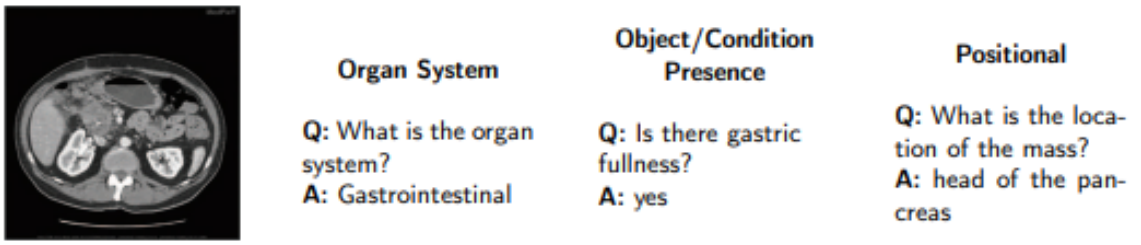


Figure 1.1: Medical VQA with VQA-RAD dataset

A cornerstone of this project involves an investigation into two cutting-edge methodologies: Meta-Learning and Contrastive Denoising Auto-Encoder (CDAE). These techniques present distinctive approaches to address the data scarcity issue, catering to the unique demands of medical VQA.

1.2 Meta Learning in Medical VQA

Meta-Learning, a paradigm within supervised machine learning, offers a powerful solution to the challenge of data scarcity by enabling rapid adaptation to new tasks with limited labeled data. The central premise of Meta-Learning is to train a model's parameters in a way that facilitates quick adaptation across a range of related tasks. In the context of Medical Visual Question Answering (VQA), where annotated data is often limited, MAML proves to be particularly valuable.
MAML's methodology involves a two-step process: pre-training and fine-tuning. During pre-training, the model is exposed to a variety of tasks and learns a set of initial parameters that are adaptable. Fine-tuning involves updating these parameters with a small amount of labeled data specific to the target task, allowing the model to swiftly adapt to the new task's nuances.

In the context of medical VQA, MAML's ability to generalize from a limited labeled dataset is crucial. The scarcity of annotated medical images and questions necessitates an approach that maximizes the utility of available data. MAML's flexibility to adapt and learn from few-shot examples ensures that the model can effectively comprehend and answer a wide array of medical questions, even with sparse training data.

1.3 Unsupervied Learning in Medical VQA

The Contrastive Denoising Auto-Encoder (CDAE) presents an innovative unsupervised learning approach that addresses the scarcity of labeled data by effectively utilizing an abundance of unlabeled medical images. In the landscape of Medical Visual Question

Answering (VQA), where annotated data is limited and expensive to acquire, CDAE's methodology holds significant potential.

CDAE operates by reconstructing clean images from noisy inputs. It introduces noise to the input images and then attempts to recover the original, noise-free versions. This denoising process forces the model to capture essential features and patterns present in the data, effectively improving representation learning without the need for explicit labels.

In the context of medical VQA, CDAE offers a dual advantage. Firstly, it extracts valuable information from unlabeled medical images, aiding in building robust and comprehensive visual representations. Secondly, these enriched representations can enhance the performance of VQA systems, enabling them to answer questions more accurately and contextually.

CDAE's application is particularly relevant to the medical field, where annotated data is scarce and domain expertise is essential for accurate annotation. By leveraging the power of unlabeled data, CDAE presents a means to navigate the challenges of data scarcity, ultimately contributing to the development of more effective and reliable medical VQA systems.

In the following sections, we will delve deeper into CDAE's mechanics, explore its potential benefits within the medical VQA framework, and discuss its implications for addressing the limited availability of properly annotated and categorized training data.

1.4 Motivation

Traditional approaches to medical image interpretation often rely on manual examination by highly skilled radiologists and medical professionals. While their expertise is invaluable, this process can be time-consuming, subjective, and prone to human errors. Moreover, in the face of a global shortage of specialized healthcare professionals, it is becoming increasingly challenging to meet the growing demand for image analysis.

There are a large number of unlabeled medical images available in a repository. These images belong to the same category as the medical Visual Question Answering image

dataset. Consequently, if an unsupervised DL model is trained using these un-annotated medical images, the learned weights may be more easily adjustable to the given problem compared to using already trained weights from ImageNet image dataset.

VQA-RAD is essentially designed to train the medical VQA task. However, we can easily create advanced classes for the given data. These newly generated classes provide an opportunity to utilize modern meta-learning techniques, which Would then be used to easily adjust the weights, to tackle the challenge of VQA in the future.

2

A Review on Medical VQA

2.1 Review on Existing Medical VQA Methods

In the initial phase of the research, a major focus has been given to reviewing existing cooperative spectrum sensing methods. Model-Agnostic Meta-Learning (MAML) has emerged as a powerful approach to enable fast adaptation of deep networks in the context of few-shot learning and optimization. In 2009, Chelsea Finn et al. proposed a novel meta-learning framework that seeks to learn a model-agnostic initialization, enabling rapid adaptation to new tasks with limited labeled data. By optimizing the model's parameters with respect to a distribution of tasks during meta-training, MAML facilitates effective generalization across tasks, making it well-suited for scenarios where acquiring abundant labeled data is challenging or impractical [1].

In medical imaging tasks with limited labeled data, CDAE can benefit from expert annotations provided by medical professionals. These annotations might include precise labels for specific abnormalities or regions of interest in the medical images. In 2021, Lixin Fan et al., proposed Collaborative Denoising Autoencoder (CDAE) as a solution to enhance the image classification performance when data is scarce. The CDAE model is designed to learn meaningful image representations by incorporating collaborative information from external sources or related datasets. By leveraging this additional information, the CDAE effectively complements the limited labeled data and captures essential visual features, even with small training sets [2].

2.2 Research Analysis of Existing Methods

In 2023, João Daniel Silva et al., proposed a novel approach to address the problem, which combines a strong image encoder based on EfficientNetV2 with a multimodal encoder based on the RealFormer architecture. The model is pre-trained through a strategy that includes a contrastive objective, and the final fine-tuning to the VQA task uses a

loss function that specifically addresses class imbalance [3].

In 2022, Meiling Wang et al., proposed a novel and effective approach to Medical Visual Question Answering, combining question-type reasoning and semantic space constraint to enhance answer accuracy and system efficiency.

The proposed method encompasses a two-step process: question-type reasoning and semantic space constraint. In the first step, the question is analyzed to identify its type, thereby facilitating the selection of the most appropriate reasoning strategy.

In the second step, a semantic space constraint is applied to restrict the search space of the AI model, ensuring that it focuses only on the relevant visual and textual information to provide accurate answers. This constraint reduces computation overhead and enhances the efficiency of the MVQA system without compromising the quality of responses [4].

The paper titled "AMAM: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering" by lHaiwei Pan presents an innovative approach to address the challenges in Medical Visual Question Answering (MedVQA). The proposed model, AMAM, leverages multimodal learning techniques to combine information from textual questions and medical images, aiming to provide accurate answers to medical queries.

AMAM incorporates advanced attention mechanisms to dynamically focus on informative regions in both the image and question, enabling it to capture intricate visual and textual cues simultaneously. Through this approach, the model establishes intricate correlations between medical images and corresponding textual queries, enhancing the overall performance of MedVQA systems [5]

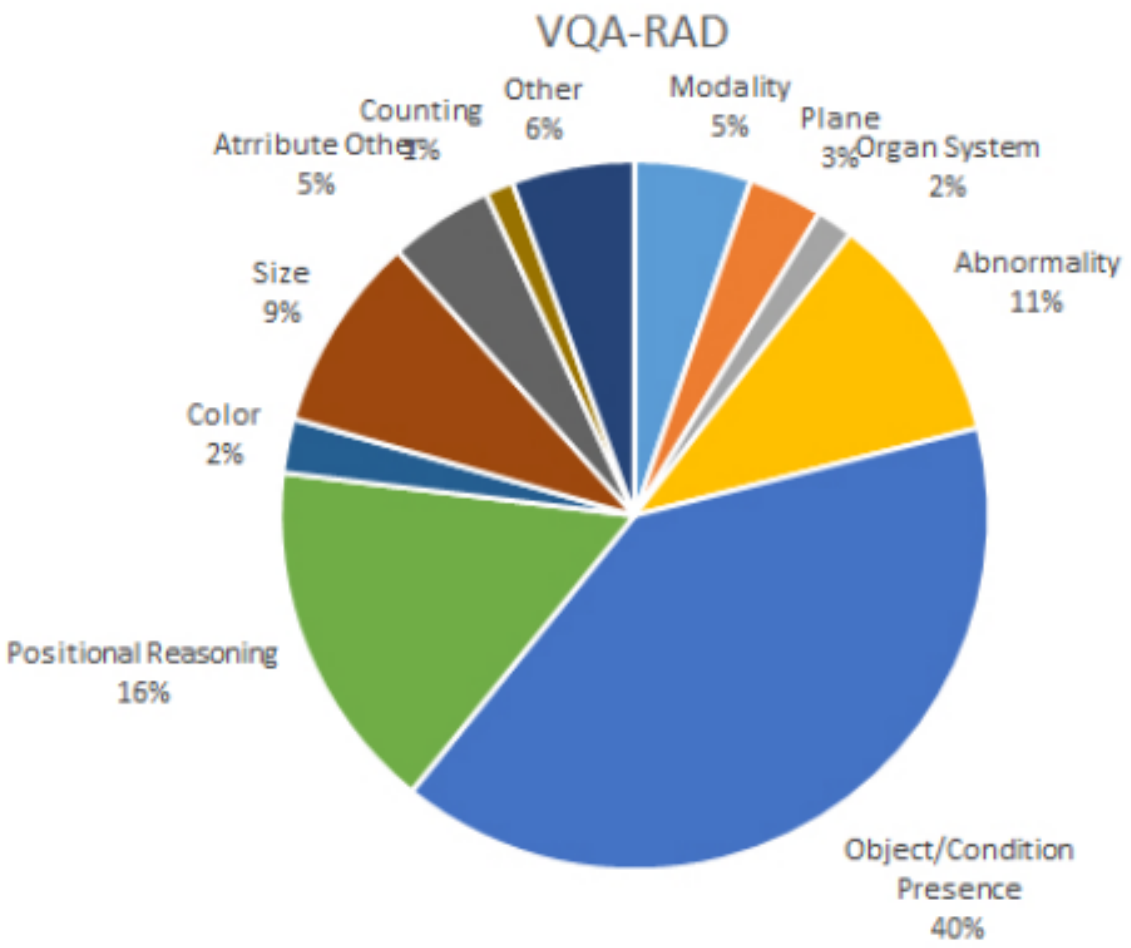


Figure 2.1: The category distribution of questions in VQA-RAD

2.3 Research Gaps

One of the major issues in medical Visual Question Answering is the lack of large, diversified, and well-annotated datasets. Medical data is often sensitive and challenging to gather, which limits the availability of publicly accessible datasets for training and evaluating models.

Accurate and reliable annotations for medical images and questions require expertise from medical professionals. However, gathering annotations from experts can be time-consuming and expensive, making it difficult to have a well-annotated dataset for training.

Medical questions can be highly specialized and domain-specific, requiring a deep understanding of medical concepts, terminologies, and context. Existing VQA models may struggle to answer complex medical questions effectively.

Medical VQA involves processing both visual and textual information. Combining these modalities effectively and learning meaningful representations from them is a complex task.

In medical practice, quick responses are often required in critical situations. Achieving real-time or near-real-time inference in medical VQA systems is a challenge, especially when dealing with large-scale medical data.

The capability of the existing frameworks to represent professional knowledge contained in the clinical questions and fine-grained details in the medical images is not enough.

3

Problem Statement based on Identified Research Gaps

3.1 Problem Formulation

In recent years, the field of medical Visual Question Answering (VQA) has gained substantial attention due to its potential in aiding medical professionals and researchers in extracting valuable insights from medical images through natural language queries. However, a major challenge in this domain is the limited availability of annotated data. The scarcity of labeled data hinders the development and deployment of accurate and reliable VQA systems for medical images.

The scarcity of data poses a critical hindrance to the development of robust and precise VQA models tailored for the medical domain. The task at hand requires a comprehensive understanding of the challenges posed by this dearth of data, ranging from the accurate representation of medical images to the effective handling of diverse and complex queries posed by medical professionals.

The primary objective of this research is to address the aforementioned challenge of data scarcity in the context of medical Visual Question Answering. Our aim is to develop a robust and effective framework that enables the VQA system to perform exceptionally well even with a limited amount of annotated medical image and question pairs. By doing so, we seek to enhance the practical applicability and reliability of VQA systems in real-world medical scenarios.

The proposed framework will leverage innovative techniques from the field of meta-learning and unsupervised learning to enable the model to learn from a small number of examples and generalize effectively to unseen medical VQA tasks. By learning from diverse medical images and questions, our framework will be designed to extract and encapsulate domain-specific knowledge that is crucial for accurate answering of medical queries.

Furthermore, a comparative analysis will be conducted to evaluate the performance of our framework against existing approaches for medical VQA.

3.2 Thesis Objective

The objectives of this thesis are as mentioned below :

- To develop a framework that effectively tackles the challenge of limited data availability in the context of the medical Visual Question Answering (VQA) problem.
- To conduct a comparative analysis between the framework and existing approaches, showcasing the advantages of our solution in addressing the data scarcity issue and achieving superior results in the medical VQA domain.

4

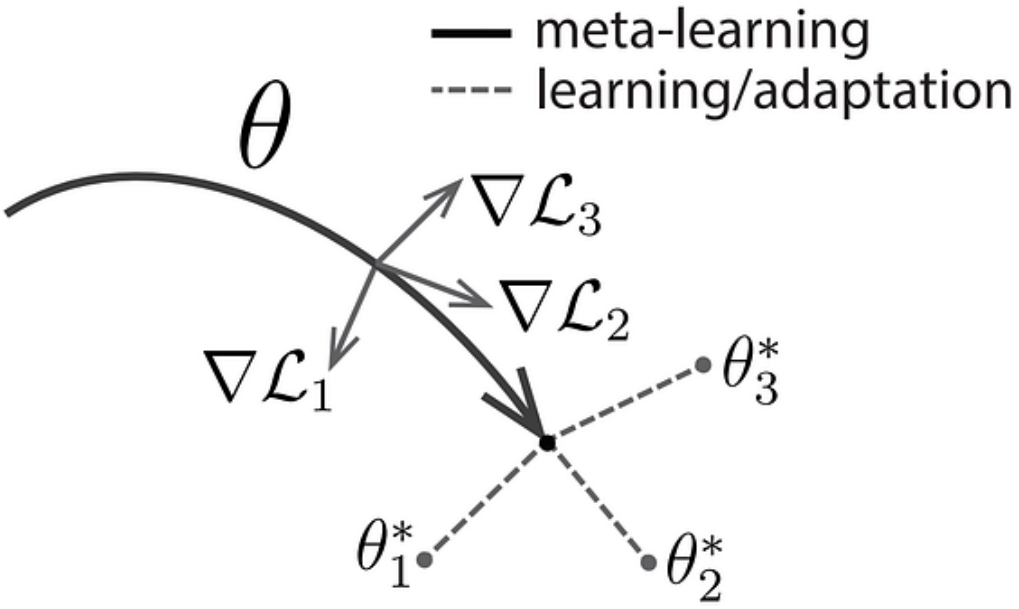
Proposed Methodology

4.1 Model-Agnostic Meta-Learning

In the context of enhancing the efficiency and adaptability of our Medical Visual Question Answering (VQA) model, we turn to the paradigm of Model-Agnostic Meta-Learning (MAML). MAML is a powerful framework designed to expedite the adaptation of a model to new tasks using minimal task-specific training data. This methodology aligns seamlessly with the challenges presented by medical VQA, where the availability of annotated data for each specific medical scenario can be limited.

Objective of MAML for Medical VQA: The core objective of incorporating MAML into our Medical VQA model is to establish an initial set of model parameters that enable swift adaptation to diverse medical tasks. Rather than training our model separately for each medical scenario, MAML empowers our model to rapidly fine-tune itself to new tasks with only a few gradient updates. This is achieved by training the model across a spectrum of medical tasks, learning a flexible parameter initialization that supports prompt adaptation.

Enhancing Medical VQA Adaptability: By employing MAML, our Medical VQA model acquires the capability to efficiently adapt to novel medical tasks. Through the meta-training process, MAML identifies model parameters that facilitate rapid fine-tuning, ensuring that our model can generalize well to a diverse range of medical scenarios.



For every task i=1,2,3, the parameter θ should reach a near-optimal parameter θ. Illustration from [9].*

- Every task $\mathcal{T}_i$ has an optimal parameter θ_i^*.
- For every task, the adaptation along the gradient $\nabla\mathcal{L}_i$ provides a parameter $\theta_i' := \theta - \alpha\nabla\mathcal{L}_i$ that should be close to θ_i^*.

Figure 4.1: How MAML should work at meta-test time

4.2 Collaborative Denoising Autoencoder

In the field of enhancing our Medical Visual Question Answering (VQA) model, we turn our attention to the innovative Collaborative Denoising Autoencoder (CDAE) technique. CDAE stands as a robust approach for jointly addressing the challenges of noise reduction and feature learning in data, a context that resonates profoundly with the complexities of medical image-based question answering.

Purpose of CDAE in Medical VQA: The fundamental aim of integrating the Collaborative Denoising Autoencoder (CDAE) into our Medical VQA framework is to enhance the quality of input data by simultaneously reducing noise and extracting meaningful features. In medical imaging, where the presence of noise can obscure critical details, and

relevant features are pivotal for accurate diagnosis, CDAE assumes a role of significance.

CDAE operates with the following core attributes:

1. Denoising Capability: CDAE is equipped with the capacity to denoise data by reconstructing clean samples from noisy inputs. This is particularly valuable in medical images, where noise reduction contributes to improved feature extraction and enhanced image quality.

2. Autoencoder Architecture: CDAE adopts the autoencoder architecture, comprising an encoder and a decoder. The encoder transforms the input into a latent representation, and the decoder reconstructs the original input from this representation.

3. Collaborative Learning: The term "collaborative" in CDAE signifies its approach to learning from multiple related inputs. This is especially beneficial for medical VQA, as it leverages collaborative insights from medical images and questions to refine feature learning.

4. Feature Learning: Beyond denoising, CDAE excels at extracting intricate features from complex data. In medical VQA, where image and text components intertwine, CDAE enhances the extraction of features that facilitate coherent image-question understanding.

By addressing noise reduction and feature learning, CDAE empowers our model to navigate the intricacies of medical data and questions, contributing to improved diagnostics, insights, and overall performance in medical image-based question answering.

4.3 System Architecture And Methodology

In the landscape of machine learning, especially with the advent of deep learning, the need for extensive labeled training data has remained a persistent challenge. Traditional machine learning algorithms and even deep learning models demand substantial amounts of labeled data to effectively learn and perform well on new tasks. This poses a significant obstacle, particularly when confronted with scenarios where data availability is limited. A groundbreaking approach to overcoming this limitation comes in the form of meta-learning.

4.3.1 Meta Learning

A complicated skill can currently be learned from scratch by AI systems, but it takes a lot of time and practice. However, we cannot afford to train every ability in every situation from scratch if we want our agents to be able to learn multiple talents and adapt to a variety of contexts. Instead of approaching each new work separately, we need our agents to learn how to learn new tasks more quickly by applying prior expertise. This is a crucial first step in creating adaptable agents that can continuously learn a wide range of activities throughout the course of their lifespan.

Diverging from the conventional machine learning paradigms, meta-learning offers a revolutionary perspective on addressing the data limitation quandary when mastering novel tasks. Unlike traditional methodologies, which require a sizable dataset even when leveraging pre-trained models from other classification problems, meta-learning focuses on the specific challenge of data scarcity during the learning process for new tasks.

In recent times, a notable advancement in the realm of meta-learning has emerged under the name of Model-Agnostic Meta-Learning (MAML). This novel approach, introduced by pioneering researchers, provides an elegant solution to the conundrum of learning from sparse data when dealing with new tasks. The fundamental premise of MAML lies in its capability to cultivate a meta-model, characterized by network weights or parameters, derived from existing tasks. This meta-model is thoughtfully crafted to exhibit a broad

suitability across an array of tasks.

The benefit of MAML's method becomes apparent when dealing with novel tasks that only require a small no. of training data. By having cultivated a versatile and adaptable meta-model, MAML-enabled models can swiftly and effectively adapt to these new tasks. The inherent efficiency of MAML stems from its unique design, which enables models to rapidly fine-tune themselves using only a small set of training images.

Within our framework, the foundational element responsible for extracting features from images is initiated through pre-trained weights originating from both MAML and CDAE. Subsequent to this initialization phase, the VQA framework embarks on an end-to-end fine-tuning process utilizing medical VQA data. In the forthcoming sections, we delve into a comprehensive exposition of the architectures governing MAML, CDAE, and the amalgamation of the two within our innovative framework.

The objective of meta-learning is to educate a model across a spectrum of distinct learning tasks, with the ultimate aim of equipping it to effectively address novel learning tasks even when presented with a limited set of training examples.

The Meta Agnostic Meta Learning Model is represented as a function f_θ with parameters θ. When the model adapts to new task T_i, the parameters θ will become θ_i'.

If the dataset for the MAML training be $D = \{x_i, y_i\}^N$.

Let number of samples = N.

Pair of Image, Class Label = $\{x_i, y_i\}$

The dataset for each MAML task is given as $D' = \{x_i', y_i'\}^{N'}$.

The samples of this dataset come from dataset of N classes, D' contains n classes. This dataset with n classes is further divided into Validation dataset and Training Dataset : D^{val} and D^{tr} respectively. In the training dataset each of the classes will contain k different images.

Training will be done as described in the following algorithm :

Algorithm 1 Overview of the meta-training procedure

procedure META-TRAIN($\mathcal{D}$, model f_θ)
  Initialize model parameters θ
  for $h = 1$ *to* H **do** ▷ Meta-update Loop
    Create meta-batch of tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_m\}$
    for each task $\mathcal{T}_i$ **do**
      Sample data $\{\mathcal{D}_i^{tr}, \mathcal{D}_i^{val}\}$ of task $\mathcal{T}_i$
      Update task models with Eq. (1) using samples from $\mathcal{D}_i^{tr}$
    Update meta-model θ with Eq. (2) using $\{\mathcal{D}_1^{val}, \mathcal{D}_2^{val}, ..., \mathcal{D}_m^{val}\}$

Figure 4.2: MAML Training Algorithm

m tasks are generated in each iteration h and a meta-batch is formed for meta learning training.

The adapted parameters θ_i' for each task T_i are calculated as follows :

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$$

The step size α may be fixed as a hyperparameter or metalearned. The model parameters are trained by optimizing for the performance of f_θ with respect to θ across tasks sampled from p(T). More concretely, the meta-objective is as follows:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)})$$

Note that the meta-optimization is performed over the model parameters θ, whereas the objective is computed using the updated model parameters θ' . In effect, our proposed method aims to optimize the model parameters such that one or a small number of gradient steps on a new task will produce maximally effective behavior on that task.

The meta-optimization across tasks is performed via stochastic gradient descent (SGD), such that the model parameters θ are updated as follows:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$$

where β is the meta step size.

The MAML model comprises 4 convolutional 3×3 layers with 2 stride, culminating in pooling (mean). Every layer is equipped with filters(64) and then a Rectified Linear Unit layer. After training, the meta-model weights will be used for fine-tuning in the VQA framework.

4.3.2 Denoising Auto-Encoder (DAE)

In field of medical domain, where scarcity of labeled datasets poses a formidable challenge, the training process for machine learning models can often be hindered by inefficiency. In light of this, the utilization of unlabeled data, which is readily accessible, emerges as a compelling strategy to enhance training efficacy. A key player in this endeavor is the Denoising Auto-Encoder (DAE), a versatile solution that unlocks the potential of unlabeled data for robust feature extraction.

At its core, an Auto-Encoder is a remarkable innovation that operates without relying on any explicit label information. Instead, it is designed to decode and subsequently reconstruct input data, thereby effectively distilling high-level features in an unsupervised manner. This intrinsic ability of Auto-Encoders aligns seamlessly with the objective of harnessing the power of unlabeled data in the medical landscape.

In our pursuit to capitalize on the substantial benefits conferred by large-scale unlabeled datasets and to bolster the model's resilience against the noise inherent in input images, a potent solution arises in the form of CDAE. By incorporating CDAE as a pivotal component which extracts features from images within our framework, we can amplify extraction of salient features from medical images. CDAE's distinctive architecture, rooted in convolutional layers, empowers it to effectively denoise input images and extract robust features despite the presence of noise.

Encoder component is responsible for mapping an input image, denoted as x, into z which is latent representation of x that preserves a substantial proportion of relevant 6 details. After that, the decoder component converts this latent representation z into the desired output y.

$$L_{rec} = \|x - y\|_2^2$$

During training, the objective of the algorithm is to reduce the error of reconstruction bw the output and the input which is y & x respectively.

4.4 Medical VQA Framework

Each input text of the question is trimmed into a twelve words sentence. Zero padding is done if the length of the sentence is less than twelve.The 300-D word embedding (GloVe) and embedding from VQA-RAD (augmenting embedding) are combined to create a 600-D vector for each word. Embedding of the question(fq), is generated by feeding the word embedding into a 1024-D LSTM.

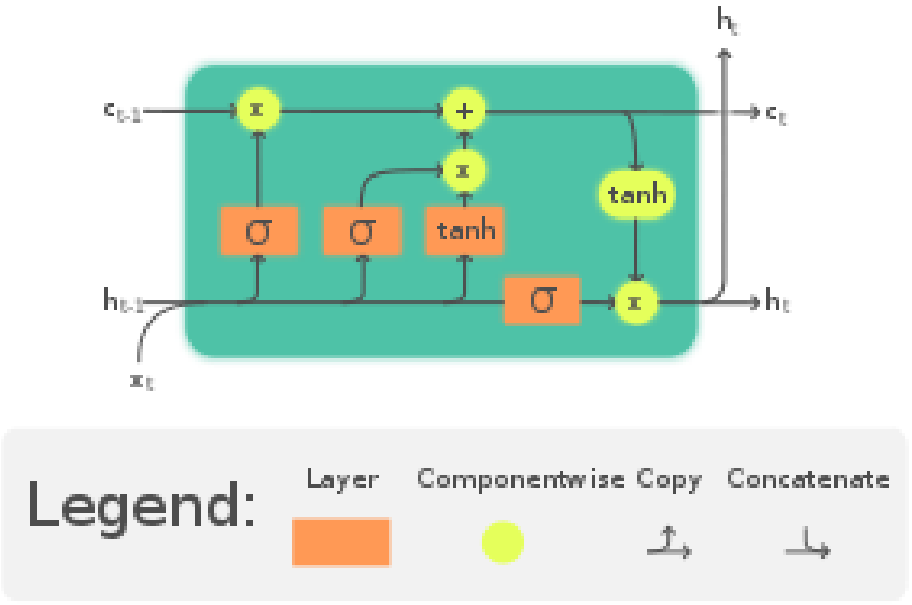


Figure 4.3: Long Short Term Memory Cell

128-D representation of the input image is formed by concatenating two 64-D vectors which are produced by the MAML and the encoder.

An attention mechanism (BAN or SAN) receives input from an image feature (fv) and a question embedding (fq) to create a joint representation (fa). A multi-class classifier accepts this feature, fa, as input. In order to integrate the successful application of the CDAE to VQA into the proposed model's training, we use a multi-task loss function.

$$L = \alpha_1 L_{vqa} + \alpha_2 L_{rec}$$

Figure 4.4: Loss Function

where Lrec indicates the loss of reconstruction of CDAE and Lvqa represents the Cross Entropy loss for VQA classification.

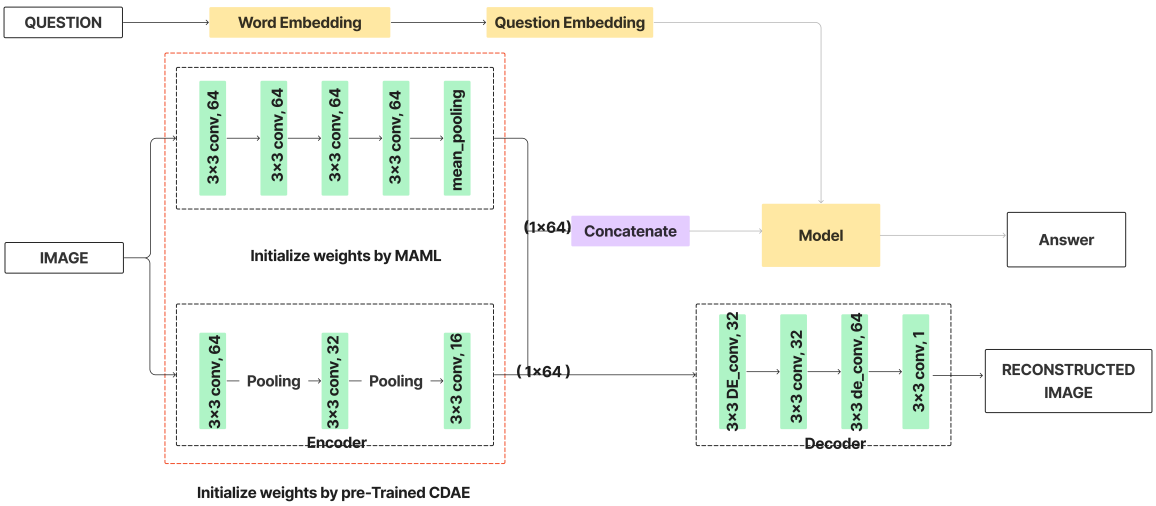


Figure 4.5: Workflow Diagram of the Framework

5

Experiment and Results

5.1 Dataset

The VQA-RAD [6] dataset consists of 3,515 questions that are associated to 315 images. Each image has multiple questions attached to it. The questions are broken down into 11 categories: "Modality", "Organ", "Other", "Plane", "Positional reasoning", "Abnormality", "Attribute", "Color", "Count", "Object/Condition Presence", and "Size". The remaining questions are for training, leaving 451 questions in the test set. The requests for information may be closed-ended, meaning they don't allow for more explanation and responses are limited to "yes/no" and additional options or open-ended inquiries, i.e. the questions are open-ended and may have more than one right answer.

. 458 responses make up the dataset. A classification over the VQA is proposed for the group of responses.

5.2 Training the Framework

Using the body component labels for the images in the dataset, the images are divided into three groups: head, chest, and abdomen. Based on the understanding of question and answer pairings matching to photographs, each bodily part's visuals are further classified into three subcategories i.e., normal images, where pathology is not there, abnormal present images, where fluid, air, mass, or tumor are present and abnormal images of organs, where organs are oversized or improperly positioned. Thus there are 9 categories present.

Five problems are sampled for each iteration of the MAML training. We choose 3 classes (from of a possible 9 classes) at random for each task. We randomly choose 6 photos for each class, of which three are to be used to update task models and 3 left are utilized to update the meta-model.

We employ 12,000 unlabeled internet images, including CT abdomen images, chest X-ray images, and brain MRI images, to train CDAE. The dataset is divided into a 9,000 image train set and a 3,000 image test set. Prior to sending the input images to the

encoder, we corrupt them using Gaussian noise.

We use the trained weights from MAML and CDAE to initialize the image feature extraction part of the VQA system. VQA-RAD dataset is then used to fine-tune the entire VQA model.

We have evaluated the accuracy of the models having only CDAE, only MAML, and both CDAE and MAML. We have also shown the accuracies with finetuning and without finetuning i.e., implemented from scratch.

Reference methods	Open-Ended Accuracy (%)	Close-Ended Accuracy(%)
VGG-16 (finetuning)	24.2	57.2
MAML (from scratch) **MAML (finetuning)**	6.5 **38.2**	68.6 **69.7**
CDAE(from scratch) **CDAE(finetuning)**	13.8 **36.7**	69.2 **70.8**
MAML+CDAE(from scratch) **MAML+CDAE (finetuning)**	15.4 **40.7**	70.9 **74.1**

Figure 5.1: Results on VQA-RAD dataset

Above table shows the results of the different VQA frameworks.Only image features extraction components are different in the frameworks. BAN is used in all of the above methods.

The findings demonstrate that, for both MAML and CDAE, pretraining followed by finetuning significantly outperforms training the framework from all scratch using just VQA RAD. Outcomes further demonstrate that MAML and CDAE outperform VGG-16 finetuning on ImageNet. Additionally, the findings demonstrate that open-ended questions (OEQ) have worse accuracy than close-ended questions (CEQ).

Additionally, the gains from fine-tuning are more significant for Open Ended type Questions. We discovered OEQ are typically more challenging to respond to than CEQ, specifically OEQ ask more detailed questions and demand lengthy responses, whereas

CEQ ask only affirmative questions ("yes/no") and typically have shorter responses. This finding suggests that the suggested image feature extraction is more advantageous for description answers that require more data from the input photos.

	SAN (baseline)	MCB (baseline)	BAN (baseline)	BAN (finetuning)
Open-ended	24.2	25.4	27.6	**43.9**
Close-ended	57.2	60.6	66.5	**75.1**

Figure 5.2: Results on VQA-RAD dataset

We contrast the framework we provide with baselines. We have outlined the outcomes of using VQA frameworks, such as the SAN [7], MCB [8], and BAN [9] frameworks, as well as fine-tuning on the VQA-RAD dataset. Although alternative attention mechanisms can be used in our framework with ease, we have presented results for our framework when BAN is used as the attention mechanism.

Comparative findings for various approaches are shown in Table 2. The baselines for extracting picture features employ VGG or ResNet models trained on ImageNet and are then fine-tuned on VQA RAD. All frameworks including ours use the same prior training models (i.e., Glove [14]) and fine-tuning on VQA-RAD for the question feature extraction.

6

Conclusions and Future Scope

6.1 Conclusions

Various past research or studies have various characteristics, advantages, disadvantages and future scope. We have proposed using meta learning algorithms and unsupervised learning algorithms for medical visual question answering. In this paper we have provided an approach to medical Visual Question Answering (VQA) by combining meta-learning using MAML and image feature extraction through a denoising auto-encoder CDAE. This methodology addresses the challenge of limited labeled training data in medical VQA. The integration of CDAE allows the framework to harness insights from a vast pool of unlabeled medical images, while MAML facilitates the rapid adaptation of meta-weights to the specific VQA task. This approach not only showcases the potency of the proposed framework but also opens avenues for enhanced performance in medical VQA by effectively leveraging both labeled and unlabeled data.

6.2 Future Scope

The successful development of a Medical Visual Question Answering (VQA) system opens up several avenues for further research and enhancement. As the field of medical imaging and artificial intelligence continues to evolve, there are several compelling directions that can be explored to extend the capabilities and impact of the VQA framework.

Currently, the VQA framework predominantly relies on images for feature extraction. Exploring the integration of other modalities, such as text reports, patient histories, and additional sensor data, can enhance the overall understanding of medical cases. This could lead to more comprehensive and accurate answers by considering a broader range of information.

Bibliography

[1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," August 2017.

[2] A. Ghafar and U. Sattar, "Convolutional autoencoder for image denoising," vol. 1, no. 1-11, 31 December 2021.

[3] J. D. Silva, B. Martins, and J. Magalhães, "Contrastive training of a multimodal encoder for medical visual question answering," vol. 18, no. 200221, May 2023.

[4] M. Wang, X. He, L. Liu, L. Qing, H. Chen, Y. Liu, and C. Ren, "Medical visual question answering based on question-type reasoning and semantic space constraint," vol. 131, no. 102346, September 2022.

[5] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, and K. Shi, "AMAM: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering," *Knowledge-Based Systems*, vol. 255, no. 109763, 14 November 2022.

[6] A. D.-F. Lau, Gayen, "A dataset of clinically generated visual questions and answers about radiology images," pp. 1–6, 2018.

[7] X. G.-D. S. A. Yang, He, "Stacked attention networks for image question answering," in *In: CVPR*, 2016, pp. 414–417.

[8] Y. R.-D. R. Fukui, Park, "Multimodal compact bilinear pooling for visual question answering and visual grounding." *EMNLP*, vol. 30, no. 4, pp. 126–136, 2016.

[9] Z. Kim, Jun, "Bilinear attention networks," 2018. [Online]. Available: https://doi.org/10.1049/iet-com.2018.5245