INFOSYS-SPRINGBOARD INTERNSHIP PROJECT DOCUMENTATION

COVERING THE INTERNSHIP UNDERTAKEN WITH

On

“X-RAY MEDICAL IMAGE CAPTIONING PROJECT”

FROM

25th MAY 2024 TO 19 JULY 2024

Submitted By

Mr. PRATIK POPAT PARBHANE

MENTOR:- SUDHEER KUMAR

TABLE OF CONTENTS

ABSTRACT

Nowadays the X-ray is the most frequently used diagnostic procedure. A radiology specialist can understand the detailed information contained in an X-ray of a chest regarding the human body’s heart and lungs. Specialists in radiology are typically tasked with reviewing chest X-rays so that patients can get the right treatment. Because a doctor may see more than 100 X-rays each day in larger cities and thorough inspection necessitates skilled doctors, obtaining a thorough medical diagnosis from such Xrays is frequently laborious and time-consuming.

This project’s goal is to demonstrate deep learning techniques for autonomously extracting clinical data from X-ray images. Deep learning approaches have been combined with algorithms to tackle this difficult challenge, with promising performance. If such reports can be generated automatically by a trained model, a lot of time and effort can be saved.

In this project, some deep learning techniques such as an encoder and decoder and a pre-trained chexnet model have been used. The task of obtaining visual features from an x-ray is performed by the chexnet model which is passed to an encoder which then sends its result to a decoder these types of techniques are mainly used in image captioning which aims to produce text from an image Here, LSTM is employed as the encoder while GRU and Bi GRU is used as the decoder. Both of these are recurrent neural networks. The generated text report is evaluated by the BLEU score.

X-Ray Medical Image Captioning Project Requirements

1. Hardware Requirements

- **Computer System**: A computer with at least 16 GB of RAM, a multi-core CPU, and an NVIDIA GPU (e.g., GTX 1080 Ti or better) to speed up the training process.
- **Storage**: At least 100 GB of free disk space to store datasets, model checkpoints, and other project-related files.

2. Software Requirements

- **Operating System**: Linux (preferred), Windows, or macOS.
- **Python**: Version 3.7 or higher.

3. Libraries and Frameworks

- **TensorFlow**: For building and training the neural network models.
- **Keras**: High-level neural networks API, running on top of TensorFlow.
- **NumPy**: For numerical computations.
- **Pandas**: For data manipulation and analysis.
- **Matplotlib/Seaborn**: For data visualization.
- **NLTK/Spacy**: For natural language processing tasks (tokenization, stemming, lemmatization).
- **OpenCV**: For image processing tasks.
- **scikit-learn**: For various machine learning utilities.

4. Data Requirements

- **X-Ray Image Dataset**: A large dataset of labeled X-ray images. Each image should have a corresponding textual description. Examples include the NIH Chest X-ray dataset or any other medical image dataset with captions.
- **Caption Dataset**: Textual descriptions (captions) for the images, preprocessed according to the text preprocessing steps outlined in the project.

5. Development Environment

- **IDE/Code Editor**: Jupyter Notebook, PyCharm, VS Code, or any other preferred development environment.
- **Version Control System**: Git for version control and collaboration.

Introduction

Medical imaging is indispensable in the current diagnostic workflows. Out of the plethora of existing imaging modalities, X-ray remains one of the most widely-used visualization methods in many hospitals around the world, because it is inexpensive and easily accessible. **Analyzing and interpreting X-ray images** is especially crucial for diagnosing and monitoring a wide range of **lung diseases**, including **pneumonia**, **pneumothorax**, and **COVID-19** complications.

Today, the generation of a free-text description based on clinical radiography results has become a convenient tool in clinical practice[5]. Having to study approximately 100 X-rays daily, radiologists are overloaded by the necessity to report their observations in writing, a tedious and time-consuming task that requires a deep domain-specific knowledge.

The typical manual annotation overload can lead to several problems, such as missed findings, inconsistent quantification, and delay of a patient's stay in the hospital, which brings increased costs for the treatment. Among all, the qualification of radiologists as far as the correct diagnosis establishing should be stated as major problems.

In the COVID-19 era, there is a higher need for robust image captioning framework. Thus, many **healthcare systems outsource the medical image analysis task**. Automatic generation of chest X-ray medical reports using deep learning can assist and accelerate the diagnosis establishing process followed by clinicians. Providing automated support for this task has the potential to ease clinical workflows and improve both care quality and standardization. For that, we propose to adapt powerful models from non-medical domain.

Medical background

Radiology is the medical discipline that uses medical imaging to diagnose and treat diseases. Today, radiology actively implements new artificial intelligence approaches

There are three types of radiologists—

1)Diagnostic radiologists

2)Interventional radiologists

3)Radiation oncologists.

➢ X-Ray Medical Image Captioning Project Business Use Case

- ➢ **Improve Diagnostic Efficiency:** Reduce the time radiologists spend on interpreting X-ray images by providing automated captions.
- ➢ **Enhance Diagnostic Accuracy:** Assist radiologists by highlighting key observations, potentially reducing human error.
- ➢ **Streamline Medical Reporting:** Automate the generation of initial medical reports, freeing up radiologists to focus on more complex cases.

Problem Statement

Current Challenges

- **Time-Consuming Process:** Radiologists spend a significant amount of time analyzing and interpreting medical images.

- **High Workload:** Increasing number of medical imaging studies leads to high workloads for radiologists, potentially impacting the quality of diagnostics.

- **Human Error:** Manual interpretation is prone to human error, which can affect patient outcomes.

Deep learning-based techniques:

Rely on the end-to-end trainable networks to extract automatic features from images and map them into meaningful text . Since deep learning-based models performed very well for many other domains, this category is the most investigated in image captioning as well.

These techniques **include encoder–decoder architectures, fully connected networks, and CNNs .**

The existing methods in the literature are basically inspired by the *show-and-tell* model **an encoder–decoder model for image captioning.**

The proposed technique is based on **a visual feature extractor (the encoder), which is usually a CNN network, and a text generator**, which is an **RNN network (the decoder).**

Semantic features are computed from medical concepts detected for medical images. **CNN networks are used for feature extraction for both visual and semantic cases whereas a multi-label classifier is implemented to detect the concepts**.

- ❖ Finally, an LSTM network is implemented for language generation where beam search is also employed for a better selection of words predicted to construct the caption.

Text Pre-Processing

We **pre-processed the captions to clean the text** and **keep only significant words**. Specifically, we tokenized each caption, **converted all characters to lowercase**, **removed stop-words**, **filtered out the punctuation, and calculated the stems of the identified words.** Two other words were added to each caption to identify the beginning (<$startstart$>) and the end (<$endend$>) of the sentence. This pre-processing process was performed using the NLTK.

The maximum length of one caption for the whole training set was 50 words, which was used to pad the remaining captions at the end. Further, embeddings were calculated from these captions to capture the semantic meaning of each sentence.

Finally, each image was represented with a **vector of size 50**, encoding the corresponding caption.

Vocabulary Construction

Now that tokens were generated from captions and concepts, we constructed the vocabulary. A dictionary of sorted unique words was constructed so that a numerical value was assigned to each word based on its order in the set.

Image preprocessing

Image preprocessing is a crucial step in preparing X-ray images for training deep learning models. It ensures that the images are in a consistent format and enhances the quality of the input data, which can significantly impact the performance of the model. The **preprocessing steps include resizing, normalization, and data augmentation.**

Resizing Images

- **Objective**: Standardize the dimensions of all images to a fixed size.
- **Method**: Resize all images to a uniform size, such as 224x224 pixels, which is a common input size for CNN models.

Normalizing Pixel Values

- **Objective**: Scale the pixel values to a range that is suitable for model training.
- **Method**: Normalize pixel values to the range [0, 1] by dividing by 255.

Data Augmentation

- **Objective**: Increase the diversity of the training data and help the model generalize better.
- **Techniques**:
  - **Rotation**: Rotate images by small angles (e.g., ±15 degrees).
  - **Flipping**: Flip images horizontally.
  - **Zooming**: Randomly zoom into images.
  - **Shifting**: Translate images horizontally or vertically.

➢ Some Information About The Machine Learning Concept

Computer Vision:-

Computer vision tasks include methods for **acquiring, processing, analyzing and understanding digital images, and extraction of highdimensional data** from the real world to produce numerical or symbolic information, e.g. in the forms of decisions. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that make sense to thought processes and can elicit appropriate action.

This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory. The scientific discipline of computer vision is concerned with the theory behind artificial systems that extract information from images.

The image data can take many forms, such as video sequences, views from multiple cameras, multi-dimensional data from a 3D scanner, or medical scanning devices. The technological discipline of computer vision seeks to apply its theories and models to the construction of computer vision systems.

CNN:-

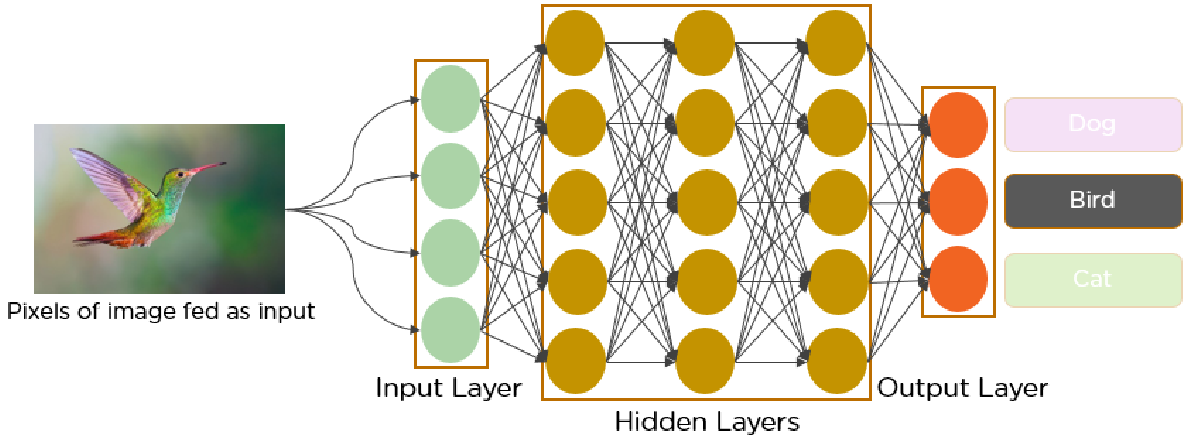
CNN stands for **Convolutional Neural Networks**. Convolutional networks are currently used in visual recognition.

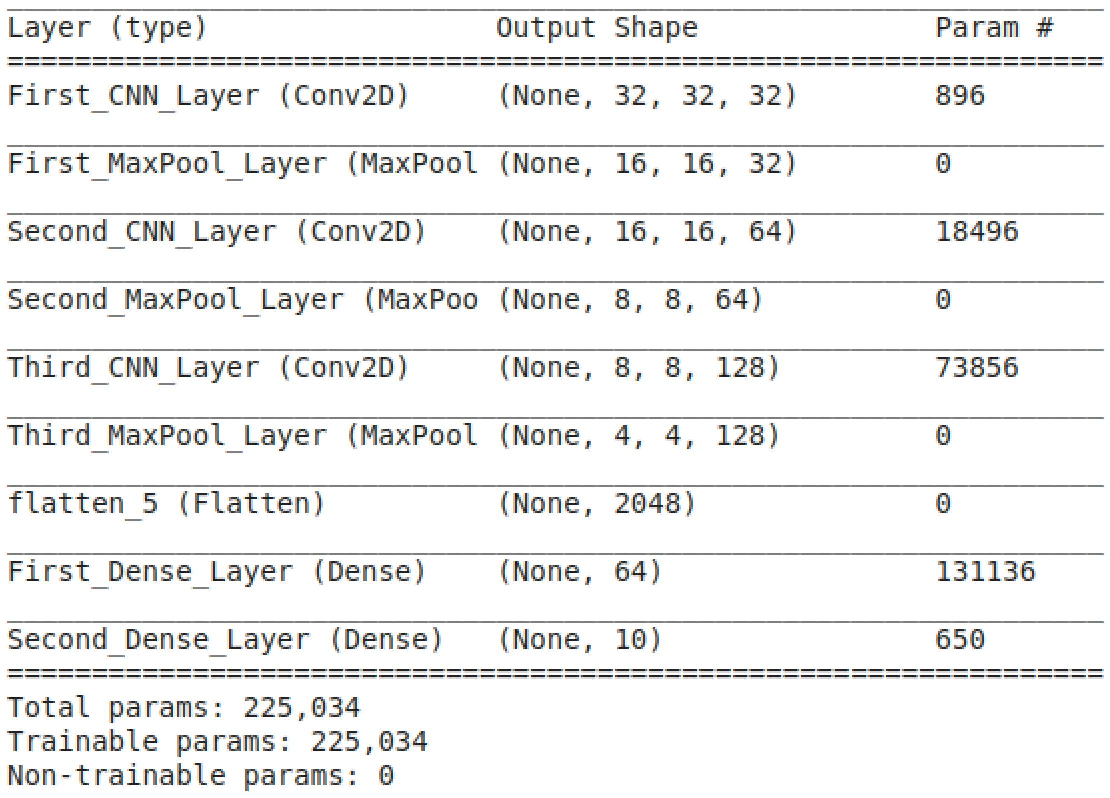
There are several convolutional layers in CNN. After these convolutional layers, text layers are fully connected layers as in a multilayer neural network.

The CNN is designed in such a way that **the benefit of the 2D structure** of the input image can be taken.

This target is accomplished with the help of several local connections and tied weights along with various pooling techniques which result in translation invariant features.

The main advantages of **using CNN are ease of training and possessing** fewer parameters as compared to other networks with an equal number of hidden states.





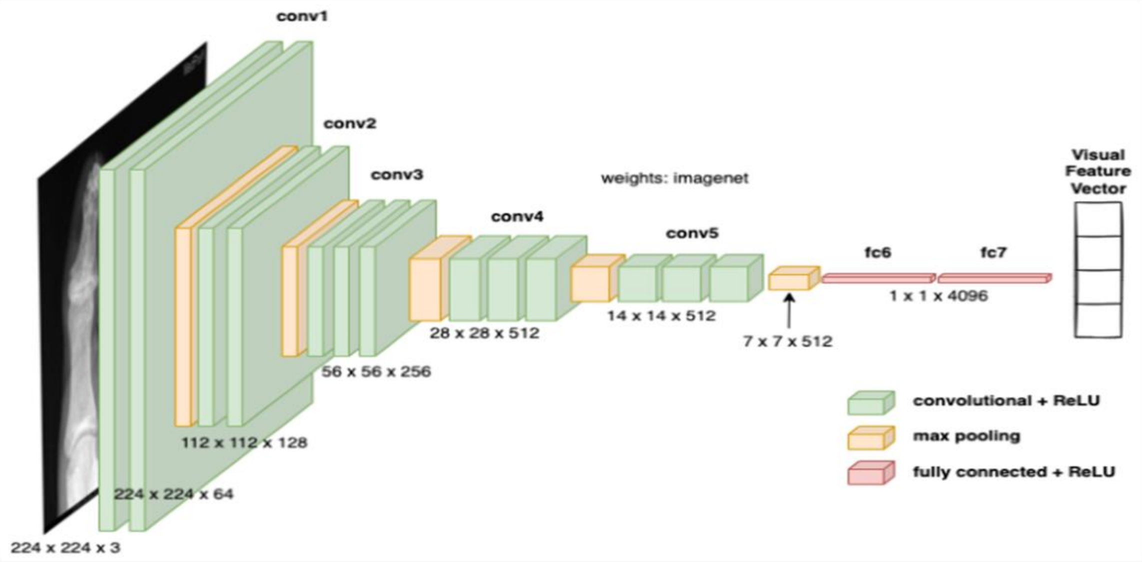
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
First_CNN_Layer (Conv2D)     (None, 32, 32, 32)        896       
_________________________________________________________________
First_MaxPool_Layer (MaxPool (None, 16, 16, 32)        0         
_________________________________________________________________
Second_CNN_Layer (Conv2D)    (None, 16, 16, 64)        18496     
_________________________________________________________________
Second_MaxPool_Layer (MaxPoo (None, 8, 8, 64)          0         
_________________________________________________________________
Third_CNN_Layer (Conv2D)     (None, 8, 8, 128)         73856     
_________________________________________________________________
Third_MaxPool_Layer (MaxPool (None, 4, 4, 128)         0         
_________________________________________________________________
flatten_5 (Flatten)          (None, 2048)              0         
_________________________________________________________________
First_Dense_Layer (Dense)    (None, 64)                131136    
_________________________________________________________________
Second_Dense_Layer (Dense)   (None, 10)                650       
=================================================================
Total params: 225,034
Trainable params: 225,034
Non-trainable params: 0
```

Visual Feature Encoding

The first step of our proposal is visual feature extraction; we employed a **pre-trained CNN model**. We used the **VGG-16 model** since it is small, has been trained on the large ImageNet dataset, and performs very well on several other classification tasks.

The model is composed of 16 layers, to which we input the medical images and generated a feature vector of size 4096 after removing the last classification layer, as illustrated in **Figure 3**. The features extracted from this layer were learned by the model while trying to predict the image class and distinguish the visual content of images. Images were pre-processed before fitting them to the VGG-16 model. They were normalized and resized to fit in the encoder and augmented with some traditional image augmentation techniques.

Figure 3. The model for visual feature extraction. The model inputs were medical images and the outputs were the visual feature vectors (size: 4096).



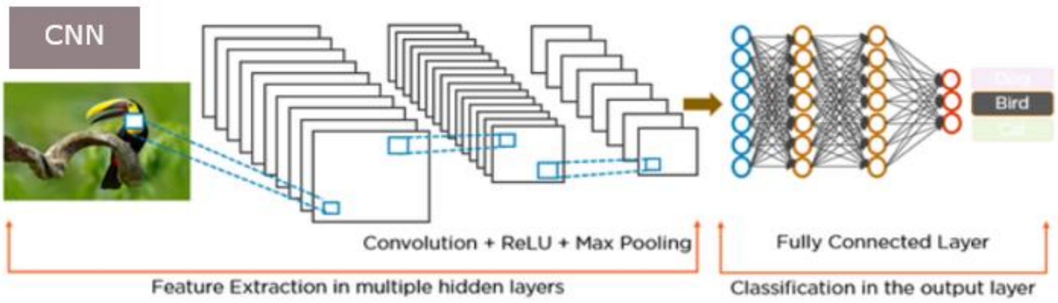
MODELS

CNN-RNN MODEL

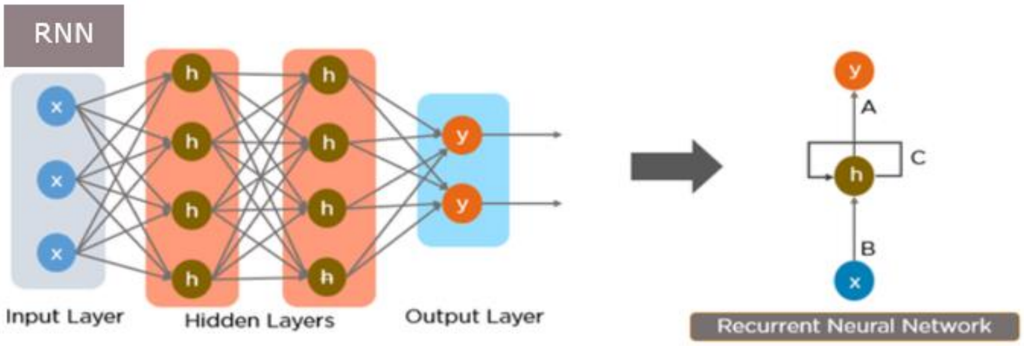
CNN-RNN framework for multilabel classification problem. The illustration of the CNN-RNN framework. It contains two parts: The CNN part extracts semantic representations from images; the RNN part models image/label relationship and label dependency.

We decompose a multi-label prediction as an ordered prediction path. For example, labels “zebra” and “elephant” can be decomposed as either (“zebra”, “elephant”) or (“elephant”, “zebra”). The probability of a prediction path can be computed by the RNN network. The image, label, and recurrent representations are projected to the same lowdimensional space to model the image-text relationship as well as the label redundancy.

Convolutional Neural Network


CNN
Dog
Bird
Cat
Convolution + ReLU + Max Pooling
Fully Connected Layer
Feature Extraction in multiple hidden layers
Classification in the output layer

Recurrent Neural Network



The main difference between RNN and CNN

The main difference between RNN and CNN come from their structure of the Neural Network. Due to their specific design, CNNs are more fit for spatial data such as images whereas RNNs are more for temporal data that comes in sequence.

CNNs employ filters within convolutional layers to transform data. Whereas, RNNs reuse activation functions from other data points in the sequence to generate the next output in a series.

CNN takes fixed size inputs and generates fixed size outputs. RNN can handle arbitrary input/output lengths.

CNN is a feed forward neural network that is generally used for Image recognition and object classification. While RNN works on the principle of saving the output of a layer and feeding this back to the input in order to predict the output of the layer.

CNN considers only the current input while RNN considers the current input and also the previously received inputs. It can memorize previous inputs due to its internal memory.

Conclusion: CNN + RNN for X-Ray Medical Image Captioning

The combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) forms a powerful architecture for the task of image captioning, particularly for medical images like X-rays. Here’s a detailed conclusion highlighting the key aspects and outcomes of using the CNN + RNN approach:

1. Image Feature Extraction with CNNs

CNNs, known for their strong ability to capture spatial hierarchies in images, serve as the backbone for feature extraction in the image captioning process. By passing the X-ray images through a pre-trained CNN (such as VGG, ResNet, or EfficientNet), we can obtain a rich set of feature maps that encapsulate the essential visual characteristics of the images. These feature maps act as a high-dimensional representation of the images, which are then used as inputs for the caption generation process.

2. Sequence Generation with RNNs

RNNs, and specifically Long Short-Term Memory (LSTM) networks, excel at handling sequential data. In the context of image captioning, RNNs are utilized to generate a sequence of words (the caption) based on the image features extracted by the CNN. The RNN processes the image features and produces one word at a time, using the previous words and the context provided by the image features to generate the next word in the sequence. This sequential prediction continues until an end-of-sequence token is generated.

3. Model Training and Optimization

The CNN + RNN model is trained end-to-end using a dataset of X-ray images and their corresponding captions. The training process involves optimizing the model to minimize the difference between the generated captions and the ground truth captions. Techniques such as teacher forcing are used during training to improve the learning process. Additionally, metrics like BLEU, METEOR, CIDEr, and ROUGE are employed to evaluate the quality of the generated captions, providing a comprehensive assessment of the model's performance.

4. Handling Medical Image Specificities

Medical images, particularly X-rays, have unique characteristics that require specialized handling. The CNN + RNN architecture is well-suited for this task due to its ability to capture intricate details in the images and generate contextually relevant captions. By fine-tuning the pre-trained CNN on medical image datasets, we can ensure that the feature extraction process is tailored to the specific nuances of X-ray images. Furthermore, the RNN can be trained with medical terminology and domain-specific language, resulting in more accurate and useful captions for medical professionals.

5. Practical Implications and Future Directions

The CNN + RNN approach for X-ray medical image captioning holds significant practical implications. It can assist radiologists and medical professionals by providing preliminary descriptions of X-ray images, potentially speeding up the diagnostic process and reducing the cognitive load on medical staff. Moreover, the automated captions can serve as supplementary documentation, enhancing the overall efficiency and accuracy of medical records.

For future work, exploring more advanced architectures, such as Transformer-based models, could further improve the quality and relevance of the generated captions. Additionally, incorporating multi-modal data, such as combining X-ray images with patient metadata, could lead to even more comprehensive and context-aware captions.

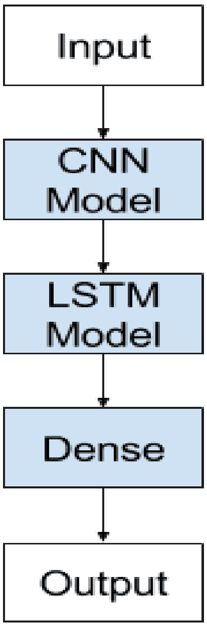
CNN-LSTM Model

A CNN-LSTM model is a combination of two powerful deep learning models: Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM). This model is often used for sequence prediction problems, particularly in the field of natural language processing, time series analysis, and more.

CNNs are primarily used for image processing, pattern recognition, and feature extraction due to their ability to effectively analyze spatial data. They are designed to automatically and adaptively learn spatial hierarchies of features from tasks with grid-like topology, such as an image.

On the other hand, LSTMs are a type of Recurrent Neural Network (RNN) that can learn long-term dependencies, making them suitable for sequence prediction tasks. They can remember or forget information from previous steps, which allows them to understand the context of the data.

When combined, a CNN-LSTM model can leverage the strengths of both architectures. The CNN part of the model can be used to extract features from the input data, while the LSTM part can use these features to make predictions based on the sequence of the data.



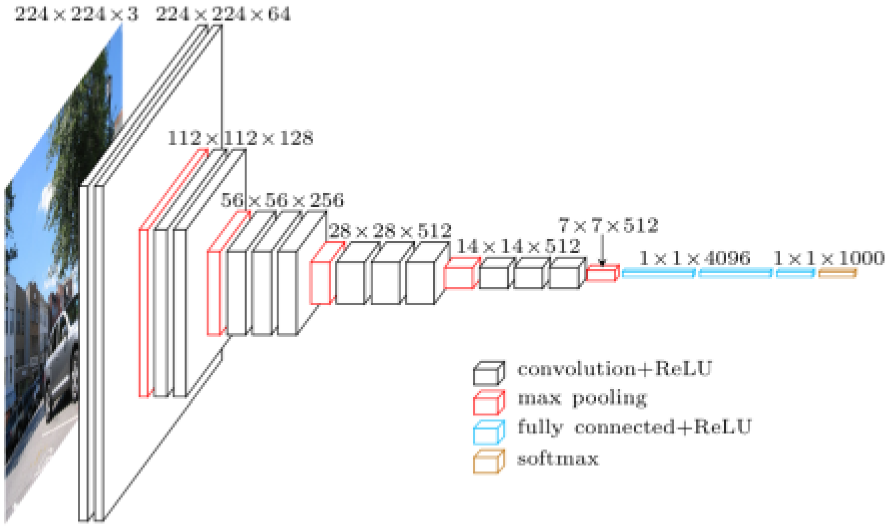
In conclusion, a CNN-LSTM model is a versatile and powerful tool for sequence prediction tasks. It can effectively extract features from the input data and use these features to understand the context and make accurate predictions. However, like all deep learning models, it requires careful tuning and large amounts of data to train effectively.

VGG16 Pre-trained

VGG16 is a pre-trained convolutional neural network (CNN) model developed by the Visual Geometry Group (VGG) at the University of Oxford. The model was introduced in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" by Karen Simonyan and Andrew Zisserman in 2014.

VGG16 is called so because it has 16 weight layers. It was trained on the ImageNet dataset, which contains over 14 million images and 1,000 classes. The input to the VGG16 network is a fixed-size 224x224 RGB image. The network consists of 13 convolutional layers with small 3x3 filters, followed by 3 fully connected layers. The final layer is a softmax layer that provides a probability distribution over the 1,000 classes.

One of the key features of VGG16 is its simplicity. Unlike some other modern architectures, VGG16 uses the same 3x3 filter size and stride of 1 throughout the network. This makes the network easier to understand and implement.

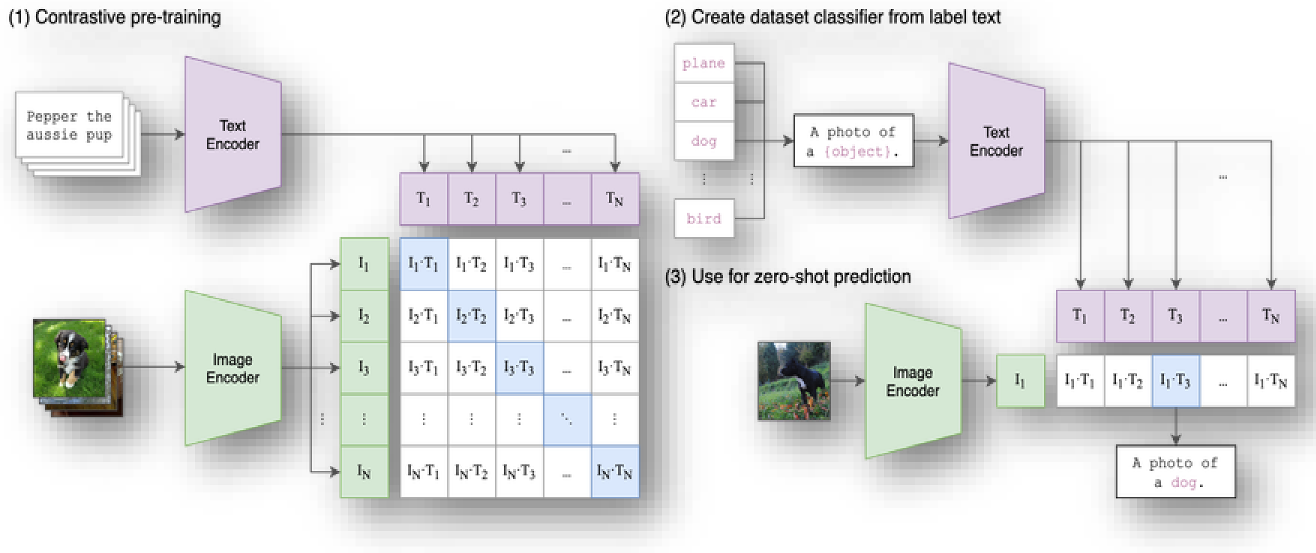


In conclusion, VGG16 is a powerful and versatile pre-trained model for image recognition tasks. Its simplicity and ability to extract high-level features make it a popular choice for transfer learning in computer vision. However, it is worth noting that VGG16 is a relatively large and computationally expensive model, which can make it unsuitable for some applications.

THIS HOW THE MODEL WORKS:-

After the model is successfully trained, we can query it with new information. (2. Zero shot)

1. Take an input
2. Encode with the custom trained encoders
3. Find a match (image or text) from the known data set.
    i. Go through each entry of the data set
    ii. Check similarity with the current input
    iii. Output the pairs resulting in most similarity
4. [Optionally] Measure the similarity between the real caption, and the guessed one.



(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
T_1 T_2 T_3 ... T_N
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

CONCLUSION

Medical image captioning, particularly for X-ray images, is a promising field that combines computer vision and natural language processing to generate automatic descriptions of medical images. This technology has the potential to assist radiologists and improve the efficiency and accuracy of diagnosis.

In recent years, significant progress has been made in X-ray medical image captioning. Deep learning models, such as CNN-LSTM architectures, have been successfully applied to extract features from X-ray images and generate corresponding captions. These models have shown promising results in generating accurate and informative descriptions of medical images.

However, there are still several challenges and opportunities for future work in this field. One major challenge is the lack of large-scale, high-quality datasets with accurate annotations. Medical data is sensitive and often difficult to obtain, which limits the size and diversity of the training data. To address this issue, future work could focus on developing methods for data augmentation and synthetic data generation.

FUTURE WORK

Another area for future work is improving the accuracy and relevance of the generated captions. Current models can sometimes generate descriptions that are factually incorrect or miss important details. To improve the accuracy of the captions, future work could explore incorporating domain-specific knowledge and developing more sophisticated language models.

Furthermore, there is a need to evaluate the performance of X-ray medical image captioning models in real-world clinical settings. Future work could focus on conducting user studies with radiologists to assess the usefulness and usability of the generated captions in clinical practice.

In conclusion, X-ray medical image captioning has the potential to revolutionize the field of radiology and improve patient care. While significant progress has been made, there are still several challenges and opportunities for future work in this exciting field.

THANK YOU