Volume: 12 Issue: IV Month of publication: April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61114

A Comparative Study of Feature Extraction Models for Image Caption Generation

Yugant Gotmare[1], Shreya Manapure[2], Tanushree Patre[3], Vatsalya Sharma[4]
Department of Artificial Intelligence, G. H. Raisoni College of Engineering Nagpur, India

Abstract: Image caption generation is a challenging task in the field of computer vision and natural language processing. This study presents a comparative analysis of various feature extraction models for image caption generation. The goal is to evaluate the performance and effectiveness of different models in capturing visual features and generating accurate and contextually relevant captions. The feature extraction models considered in this study include ResNet (Residual Neural Network), DenseNet, VGG (Visual Geometry Group), InceptionNet, DenseNet, and XceptionNet. We conduct extensive experiments on the Flickr 8k dataset, utilizing a common architecture for the caption generation task. Evaluation is performed using BLEU scores to assess the quality and similarity of generated captions to human references. Our results indicate varying performance across the different models, with notable differences in BLEU scores. We further discuss the strengths and limitations of each model, providing insights into their suitability for image caption generation. This comparative study contributes to the understanding of the impact of feature extraction models on the performance of image caption generation systems.

***Keywords*: Deep Learning, Computer Vision, Natural Language Processing**

I. INTRODUCTION

Image caption generation, the task of automatically generating descriptive captions for images, plays a vital role in bridging the gap between visual content and natural language understanding. The recent advancements in deep learning have significantly improved the performance of image captioning systems. One crucial component of such systems is the feature extraction model, responsible for encoding visual information into a form suitable for generating captions.

In this paper, we present a comparative study of different feature extraction models for image caption generation. The primary objective is to analyze and evaluate the performance of various state-of-the-art models in capturing meaningful visual features and producing coherent and contextually relevant captions. Specifically, we focus on the comparison of ResNet (Residual Neural Network), DenseNet, VGG (Visual Geometry Group), InceptionNet, DenseNet, and XceptionNet as feature extraction models. These models have demonstrated remarkable performance in various computer vision tasks and have been widely adopted in image captioning research. We conduct extensive experiments using the popular Flickr 8k dataset, consisting of a diverse collection of images with corresponding human-generated captions. A common architecture is employed to ensure a fair comparison between the different feature extraction models. To assess the quality and similarity of generated captions to human references, we utilize the BLEU (Bilingual Evaluation Understudy) metric, which measures the precision of n-gram matches between the generated and reference captions. Through our experiments and analysis, we aim to provide insights into the strengths and limitations of each feature extraction model for image caption generation. By quantitatively evaluating the performance of these models using BLEU scores, we can draw comparisons and conclude their effectiveness in capturing visual information and generating accurate and contextually relevant captions. The findings of this comparative study will contribute to the understanding of the impact of feature extraction models on the performance of image caption generation systems.

II. RELATED WORK

The study aimed to investigate how captions can improve picture understanding using transfer learning models, notably Inception V3-ARNN, VGG16-RNN, and VGG16-RNN with BLEU. The researchers used a variety of beam search techniques including bi-directional RNN while comparing various models. The I-ARNN model achieved the lowest loss of 3.0904 for the 10th epoch in the evaluation based on training and validation loss, followed by the I-RNN model at 3.1463.

After a specific number of epochs, the RNN model showed a constant validation loss, whereas the ARNN model preserved variability because of its bi-directional and time-distribution layers. Last but not least, the I-ARNN model outperformed previous InceptionV3 transfer learning models, providing more accurate captions when creating captions for testing images using a beam search parameter of k=3.

The development of recurrent neural networks (RNNs) with long short-term memory (LSTM) units for image captioning. The study shows that their method, known as CNN, performs as well as the baseline on the difficult MSCOCO dataset while requiring less time to train each parameter individually. 113,287 photos make up the training dataset, and 5,000 images each for validation and testing. The researchers also present CNN+Attn, a technique that combines attention mechanisms.

A huge number of object detectors are applied to the image to produce captions, producing several high-scoring detections. Using an intersection/union threshold of 0.3, these detections are combined into groups based on the overlapping regions. Each group in the image depicts a different thing and each group receives its object node. By using this method, it is possible to prevent the prediction of several object labels for the same area of an image, which can occur when various object detectors pick up the same thing. The researchers also scan the image with item detectors, creating nodes for the stuff categories with the best detections.

The work compares the performance of LSTM and GRU architectures for accurately describing images or surroundings such as humans using the BLEU score. The model employs a VGG16 architecture for picture feature extraction, integrating 3*3 convolution and max pooling layers. Tokenized captions are converted into dense vectors of 256 by 34 by an Embedding layer. A CNN encodes images into vector representations, while an RNN decoder generates descriptive words based on learned image attributes. This method has numerous uses in robotic vision and business.

Progress has been made as picture augmentation, picture feature extraction using VGG16, text cleaning, tokenization, and LSTM-based caption synthesis are all part of the proposed system. It ensures various and accurate captions by tokenizing the text and mapping words to integer indices. The LSTM model is then trained on the pre-processed text and visual characteristics to generate descriptive captions, and its performance is measured using the BLEU score, which measures similarity to ground-truth captions. Because of the system's adaptability, it can generate text and voice descriptions for input photos, making it useful in a variety of industries like education, research, social media, news outlets, and mobile applications.

III. METHODOLOGY

A. Dataset Description

The dataset used in this study is the Flickr 8k dataset, which serves as a benchmark collection for sentence-based image description and search. It consists of 8,000 images, each paired with five different captions, providing clear descriptions of the salient entities and events epicted in the images.

The images in the dataset were selected from six different Flickr groups, ensuring a diverse representation of scenes and situations. Notably, the dataset tends to avoid well-known people or locations, focusing on capturing a variety of everyday scenes. This characteristic enhances the generalizability of the dataset to real-world scenarios.

Preprocessing steps were performed on the dataset to prepare it for the experiments. This includes resizing the images to a consistent resolution and ensuring the captions are clear and relevant to the corresponding images. We manage to delete all digits, special characters and remove additional spaces from the captions. These preprocessing steps help maintain the integrity and consistency of the dataset for accurate evaluation and analysis.

The availability of multiple captions per image offers rich contextual information and allows for capturing diverse perspectives and descriptions. This aspect of the dataset contributes to the evaluation of the image captioning models in generating meaningful and diverse captions for a given image.

By incorporating the Flickr 8k dataset into our study, we aim to evaluate and compare the performance of various feature extraction models for image caption generation using a diverse and well-established benchmark collection.

A man in a red baseball cap eats a chip .
A boy in a hat stands barefoot at the edge of a body of water .
A man wearing a black cap and black leather coat is sitting and reading a book .
A hockey player in white tries to steal the puck from another player in red .

B. Feature Extraction Models

To extract relevant and meaningful visual features from the images, we explore several popular and state-of-the-art pre-trained deep learning models for feature extraction. These models include ResNet (Residual Neural Network), DenseNet, VGG (Visual Geometry Group), InceptionNet, DenseNet, and XceptionNet. Each of these models has demonstrated strong performance in various computer vision tasks and has been widely adopted in the research community.

We describe the architecture and characteristics of each feature extraction model in detail. These models are chosen based on their ability to capture and represent the visual information in the images effectively. By leveraging the pre-trained weights of these models, we aim to extract high-level features that can capture the salient visual attributes necessary for generating accurate and descriptive captions.

The feature extraction models are utilized to process the input images from the Flickr 8k dataset. The extracted features serve as the input for the subsequent image captioning architecture, enabling the generation of captions that align with the visual content of the images.

1) *ResNet (Residual Neural Network):* ResNet is a deep neural network architecture that has made significant advancements in various computer vision tasks. For feature extraction, we employed the ResNet50 variant, which consists of 50 layers and is pre-trained on the ImageNet dataset. The ResNet50 architecture incorporates residual connections, which enable the model to effectively address the problem of vanishing gradients and facilitate the training of very deep networks. It consists of a series of residual blocks, with each block containing multiple convolutional layers and skip connections that bypass some of the layers. These skip connections help propagate gradients more effectively and allow the model to learn both shallow and deep features simultaneously.

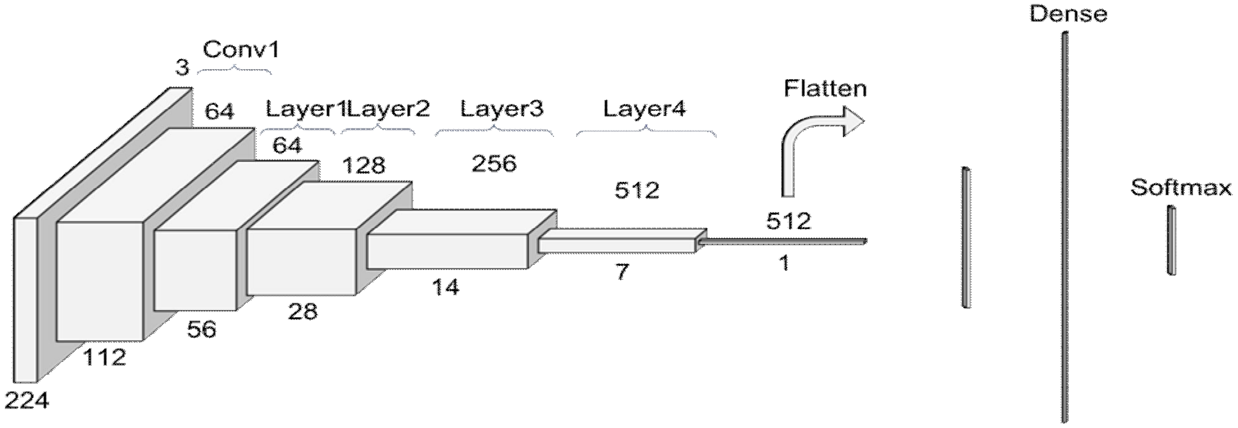


Fig 2 Architecture of ResNet

2) *DenseNet:* DenseNet is a convolutional neural network architecture known for its dense connectivity pattern, where each layer is connected to every other layer in a feed-forward fashion. Our study used the DenseNet201 variant, which consists of 201 layers and is pre-trained on the ImageNet dataset. The DenseNet201 architecture is characterized by dense blocks, which contain multiple layers that are densely connected. Within each dense block, feature maps from previous layers are concatenated together, promoting feature reuse and enhancing the flow of information through the network. This dense connectivity pattern facilitates feature propagation and enables the network to effectively capture both local and global image features.

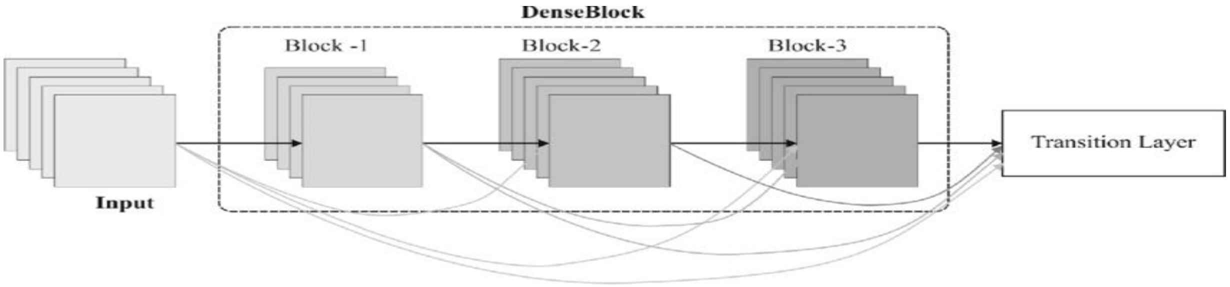


Fig 3 Architecture of DenseNet

3) *VGG16:* VGG16 is a widely used convolutional neural network architecture known for its depth and simplicity. It consists of 16 layers, including convolutional layers, max-pooling layers, and fully connected layers. In our study, we employed the VGG16 variant, which is pre-trained on the ImageNet dataset. The VGG16 architecture is characterized by a series of convolutional layers with small 3x3 filters, followed by max-pooling layers for spatial downsampling. This pattern of repeated convolutions and pooling helps capture both local and global image features at different scales. The architecture's simplicity and uniform structure make it easy to understand and interpret.

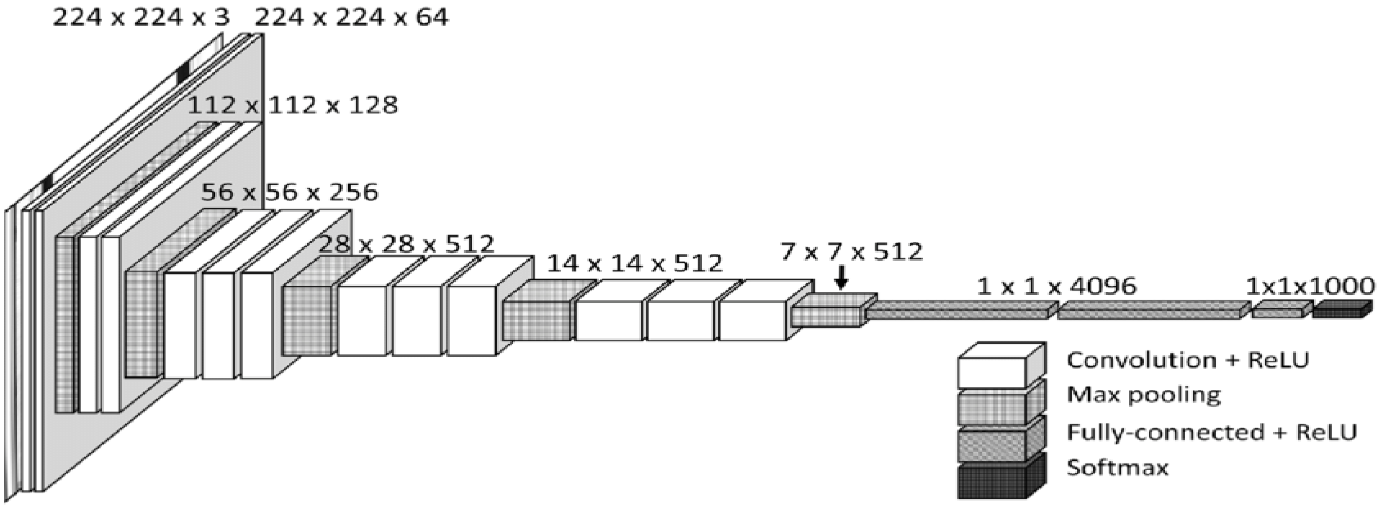


Fig 4 Architecture of VGG16

4) *InceptionV3:* InceptionV3 is a convolutional neural network architecture designed for image classification tasks. It is known for its effective use of inception modules, which enable the network to capture features at different spatial scales. The architecture is deeper and more complex compared to previous models like VGG16 and ResNet. The InceptionV3 architecture is characterized by the extensive use of inception modules, which consist of parallel convolutional operations with different filter sizes and pooling operations. This design allows the network to capture features at different spatial scales and enables it to handle variations in object sizes and shapes. The architecture's depth and multi-scale feature extraction capability make it suitable for various computer vision tasks.

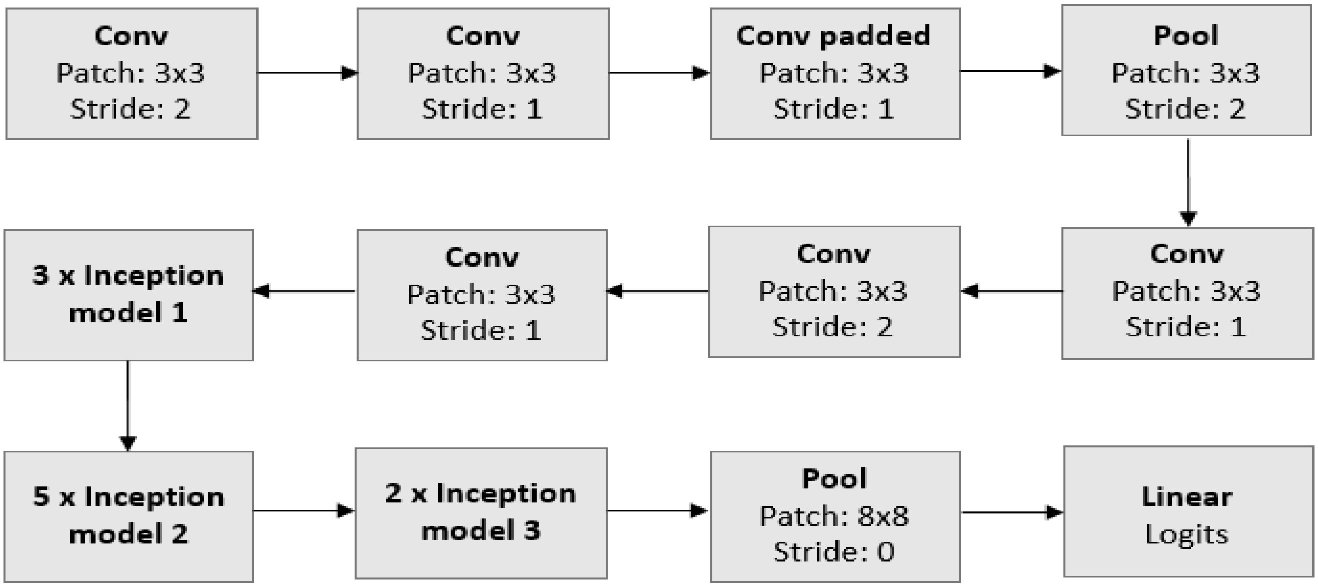

Conv Patch: 3x3 Stride: 2
Conv Patch: 3x3 Stride: 1
Conv padded Patch: 3x3 Stride: 1
Pool Patch: 3x3 Stride: 2
Conv Patch: 3x3 Stride: 1
Conv Patch: 3x3 Stride: 2
Conv Patch: 3x3 Stride: 1
3 x Inception model 1
5 x Inception model 2
2 x Inception model 3
Pool Patch: 8x8 Stride: 0
Linear Logits

Fig 5 Architecture of InceptionV3

5) *XceptionNet:* Xception is a convolutional neural network architecture that builds upon the Inception architecture. It was introduced as an extension to address the limitations of traditional Inception modules and achieve better performance in terms of both accuracy and computational efficiency. The Xception architecture employs an extreme version of depthwise separable convolutions, which decouple the spatial and channel-wise filtering operations. This design significantly reduces the number of parameters and computations compared to traditional convolutions, leading to improved efficiency without sacrificing performance. The Xception architecture is characterized by its deep and efficient feature extraction capabilities. It leverages depthwise separable convolutions to capture spatial and channel-wise dependencies efficiently. The decoupling of these operations allows the network to extract meaningful features while reducing computational complexity.

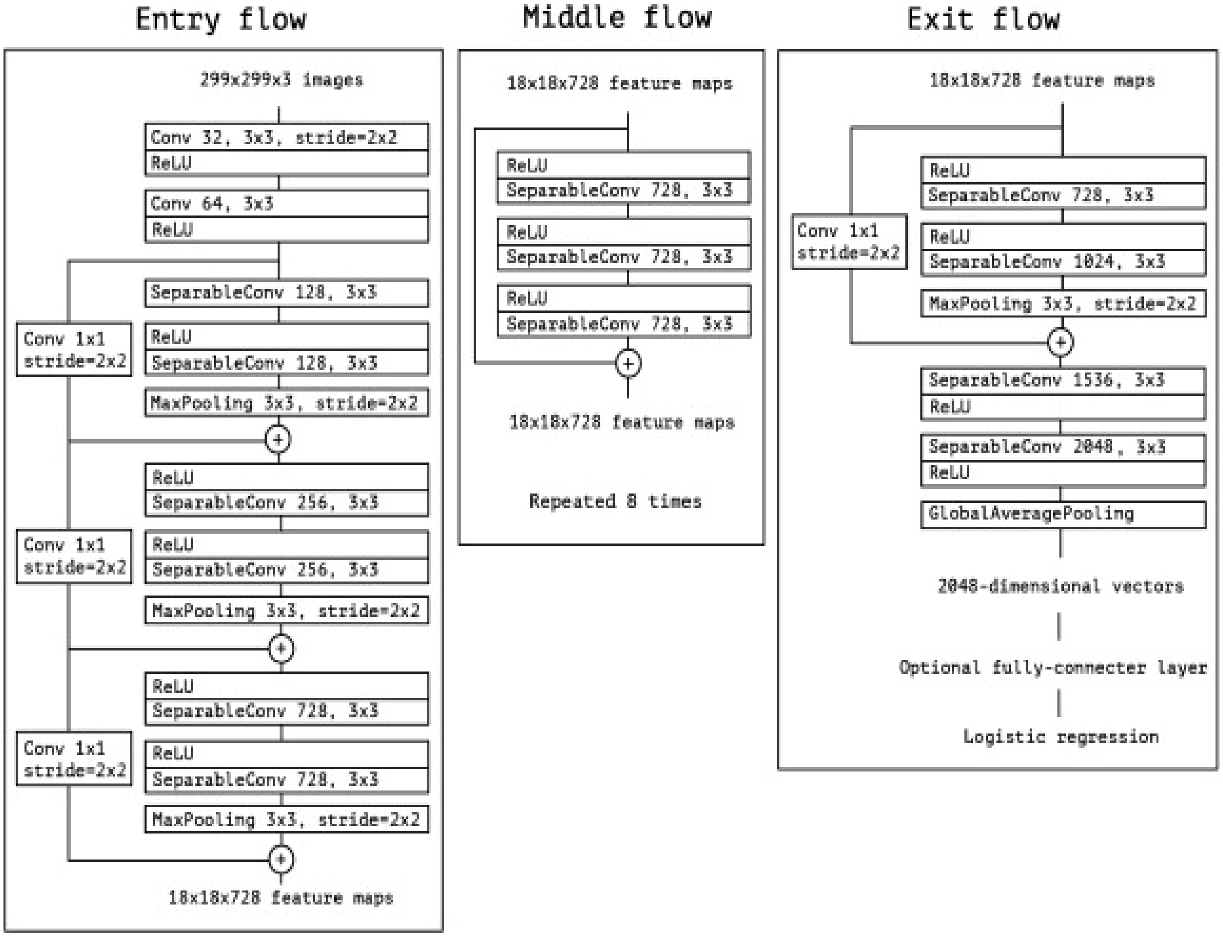


Fig 6 Architecture of XceptionNet

C. Main Architecture

The image captioning architecture comprises two main components: image feature extraction and caption generation. The image feature extraction part involves using pre-trained networks like VGG, ResNet, InceptionNet, DenseNet, or XceptionNet to extract high-level features from the input image. These features are then processed through dropout and fully connected layers to reduce dimensionality. On the other hand, the caption generation part takes a sequence of word indices representing the caption as input. The word indices are embedded into dense vectors, and dropout layers are applied for regularization. The embedded sequences are passed through an LSTM layer to capture sequential dependencies. Dropout and fully connected layers further process the LSTM output. The image features and caption features are combined using element-wise addition, followed by another fully connected layer. Finally, the model predicts the next word in the caption using a softmax-activated output layer. This architecture effectively combines visual information from the image with contextual understanding captured by the LSTM network to generate accurate and meaningful captions for input images.

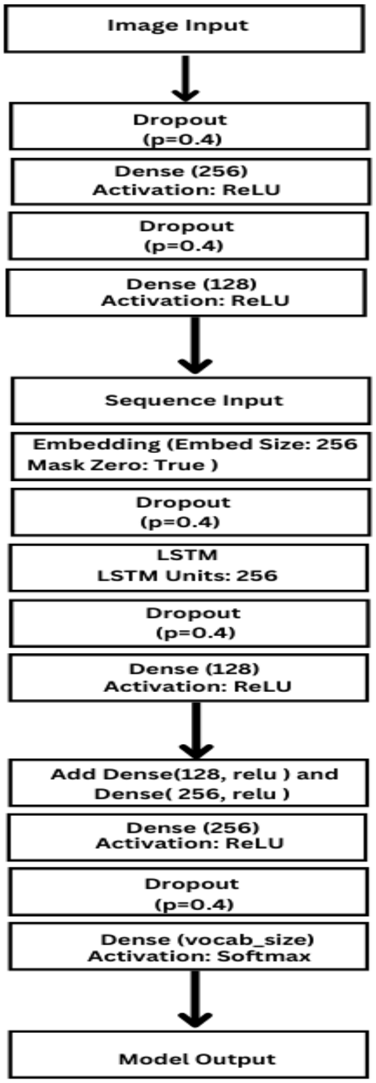


Image Input
Dropout (p=0.4)
Dense (256) Activation: ReLU
Dropout (p=0.4)
Dense (128) Activation: ReLU
Sequence Input
Embedding (Embed Size: 256 Mask Zero: True)
Dropout (p=0.4)
LSTM LSTM Units: 256
Dropout (p=0.4)
Dense (128) Activation: ReLU
Add Dense(128, relu) and Dense(256, relu)
Dense (256) Activation: ReLU
Dropout (p=0.4)
Dense (vocab_size) Activation: Softmax
Model Output

Fig 7 Model Architecture

D. *Performance Measures*

In our evaluation process, we employ the BLEU (Bilingual Evaluation Understudy) metric to assess the quality of the generated captions. BLEU is a widely used metric in natural language processing tasks, including machine translation and text generation. The value of the BLEU score always lies between 0-1.

For evaluation, we focus on BLEU-1 and BLEU-2. BLEU-1 calculates the precision of unigrams (individual words), and BLEU-2 calculates the precision of bigrams (two-word sequences). Both scores are calculated by counting the number of matching n-grams in the predicted captions and dividing it by the total number of predicted n-grams.

Precision evaluates the quality of generated captions. It measures how many of the predicted n-grams are present in the reference captions. A higher precision indicates a better match between the predicted and reference captions.

The weights parameter in the corpus_bleu function allows us to assign different weights to the precision scores of different n-gram orders. For example, weights=(1.0, 0, 0, 0) is used for BLEU-1, meaning that only the precision of unigrams is considered. Similarly, weights=(0.5, 0.5, 0, 0) is used for BLEU-2, indicating equal weightage to the precision of both unigrams and bigrams.

BLEU-1 = precision_1 = (number of matching unigrams) / (number of predicted unigrams)

BLEU-2 = precision_2 = (number of matching bigrams) / (number of predicted bigrams)

IV. RESULTS

The performance of various feature extraction models, including Inception, ResNet, VGG, DenseNet201, and Xception, for image caption generation evaluation, is based on two important BLEU metrics: BLEU-1 and BLEU-2.

The BLEU score for these models is shown in the table. The table shows that VGG achieved the highest BLEU-1 and BLEU-2 scores, indicating its superior performance in generating captions with higher word-level accuracy and better sequence consistency. ResNet and DenseNet201 also demonstrated competitive performance. On the other hand, Inception and Xception obtained lower scores, suggesting some limitations in their caption generation capabilities.

BLEU scores provide quantitative measures of caption quality, they may not fully capture the richness and creativity of human-generated captions. Thus, conducting qualitative evaluations and analyzing a diverse set of generated captions can offer valuable insights into the overall performance and behavior of the models.

V. TABLE CAPTIONS

Models	Metrics

VGG16 + Main LSTM model
Epochs = 15
Batch Size = 32
Optimizer = Adam
Loss = Categorical_Crossentropy
	BLEU-1: 0.632869
	BLEU-2: 0.414839

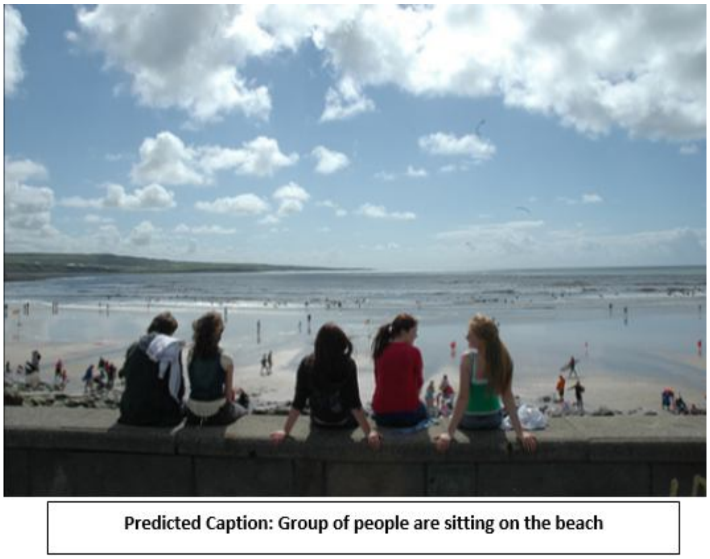
Resnet + Main LSTM model
Epochs = 15
Batch Size = 32
Optimizer = Adam
Loss = Categorical_Crossentropy
	BLEU-1: 0.597620
	BLEU-2: 0.368048

DenseNet201 + Main LSTM model
Epochs = 15
Batch Size = 32
Optimizer = Adam
Loss = Categorical_Crossentropy
	BLEU-1: 0.575660
	BLEU-2: 0.343448

InceptionV3 + Main LSTM model
Epochs = 15
Batch Size = 32
Optimizer = Adam
Loss = Categorical_Crossentropy
	BLEU-1: 0.557422
	BLEU-2: 0.331443

Xception + Main LSTM model
Epochs = 15
Batch Size = 32
Optimizer = Adam
Loss = Categorical_Crossentropy
	BLEU-1: 0.485566
	BLEU-2: 0.217358

Table No. 1: Comparison between Different Models



Predicted Caption: Group of people are sitting on the beach

Fig 8 Output

VI. CONCLUSIONS

In this study, we conducted a comprehensive analysis and comparison of various feature extraction models for image caption generation. We explored five prominent models, including Inception, Resnet, VGG, DenseNet201, and Xception, to extract image features that serve as input to our image captioning architecture. The results revealed that Vgg achieved the highest BLEU-1 and BLEU-2 scores, indicating its superior performance in generating accurate and contextually relevant captions for images. Resnet and DenseNet201 also demonstrated competitive performance, while Inception and Xception exhibited slightly lower scores. The experiments highlight the importance of selecting appropriate feature extraction models to improve the image captioning process.

While this study provides valuable insights into the effectiveness of various feature extraction models for image caption generation, there are several avenues for future research. Firstly, exploring novel techniques for feature extraction, such as attention mechanisms, could further enhance the quality of image representations. Investigating transfer learning and fine-tuning strategies specific to image captioning tasks may yield improved performance across different datasets. Furthermore, incorporating advanced natural language processing techniques and language models like BERT could lead to more contextually rich and diverse captions. Moreover, considering the impact of different evaluation metrics and exploring ensemble techniques could provide a more comprehensive understanding of model performance.

REFERENCES

[1] Takkar, Sahil, Anshul Jain, and Piyush Adlakha. "Comparative study of different image captioning models." 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2021

[2] Katiyar, Sulabh, and Samir Kumar Borgohain. "Comparative evaluation of CNN architectures for image caption generation." arXiv preprint arXiv:2102.11506 (2021).

[3] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing."Convolutional image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] Chen, Xinlei, and C. Lawrence Zitnick. "Learning a recurrent visual representation for image caption generation." arXiv preprint arXiv:1411.5654 (2014).

[5] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[6] Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." IEEE transactions on pattern analysis and machine intelligence 35.12 (2013): 2891-2903.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989

[8] Sharma, Grishma, et al. "Visual image caption generator using deep learning." 2nd international conference on advances in Science & Technology (ICAST). 2019.

[9] Kinghorn, Philip, Li Zhang, and Ling Shao. "A region-based image caption generator with refined descriptions." Neurocomputing 272 (2018): 416-424

[10] Indumathi, N., et al. "Apply Deep Learning-based CNN and LSTM for Visual Image Caption Generator." 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2023.