[1] Supriya Kurlekar
[2] Dr. Manasi R. Dixit

Exploring DenseNet for Image Captioning

Journal of Electrical Systems

Abstract: - Captioning images is a complicated process in computer vision that necessitates a combination of visual comprehension and language processing. Image captioning is highly useful in various fields like accessibility, robotics, and autonomous systems as it generates text descriptions from images automatically. Lately, DenseNet-121, a densely connected convolutional neural network (CNN), has shown great performance in image classification and transfer learning tasks. This study explores using DenseNet-121 as a feature extraction backbone in an image captioning model. We assess how well it performs in relation to other CNN models like ResNet and VGG in terms of both caption quality and computational efficiency.

Keywords: Image Captioning, DenseNet

I. INTRODUCTION

The goal of image captioning is to automatically create descriptive sentences for images by combining visual recognition and natural language generation. This task in both domains has attracted attention in areas like accessibility, content creation, and autonomous systems. Huang et al. (2017) presented DenseNet-121, a neural network with dense connections, which has demonstrated great potential in image classification by minimizing computational redundancy and promoting feature reuse. The objective of this study is to evaluate how well DenseNet-121 performs in generating image captions, focusing on the accuracy of the captions produced and comparing it with other widely used architectures.

II. RELATED WORK

2.1 Image Captioning Models

Older image captioning models included manually designed features and basic statistical language models. The rise of deep learning has made neural models like convolutional and recurrent neural networks increasingly popular. Modern image captioning systems often rely on Convolutional Neural Networks (CNNs) to extract features and use recurrent neural networks like Long Short-Term Memory (LSTM) networks for generating sentences. The effectiveness of CNN-RNN structures in generating high-quality captions was shown in models such as Show and Tell (Vinyals et al., 2015) and Show, Attend and Tell (Xu et al., 2015).

2.2 Feature Extraction in Image Captioning

Deep CNNs are highly advantageous for visual feature extraction in image captioning models. Previous methods used VGG and ResNet structures; but recent studies have explored models with complex connection patterns like DenseNet. DenseNet-121, which makes use of dense connectivity and growth rates to enhance feature reuse, has the potential to offer more diverse feature representations that may enhance the quality of produced captions.

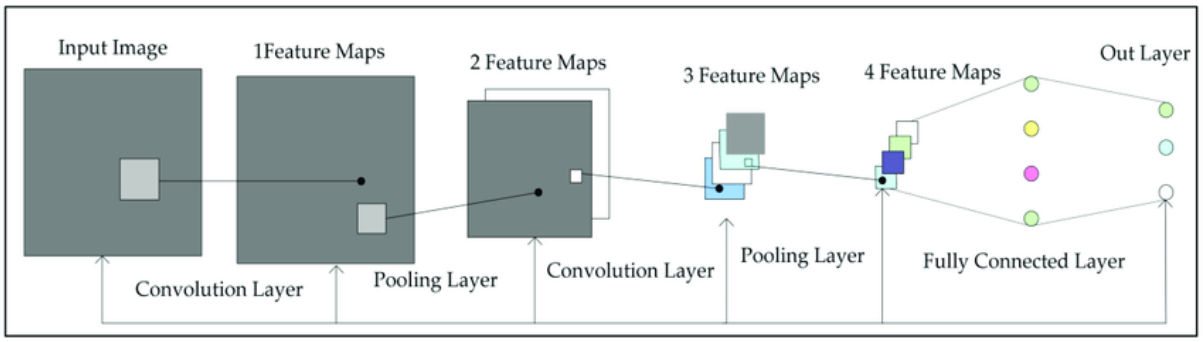

Fig.1 Feature extraction layer model

Nevertheless, when the CNN has more layers, specifically when they become deeper, the issue of 'vanishing gradient' occurs. As the path lengthens from input to output layers, some information may disappear, hindering the network's training efficiency. DenseNets address this issue by altering the CNN structure and streamlining connections between layers. In a DenseNet structure, every layer is linked directly to all other layers, which is why

[1] Assistant Professor, Department of Electronics and Telecommunication Engineering, JSPM NTC, Pune(Maharashtra), India.
Email: supriyakurlekarsit@gmail.com
[2] Professor, Department of Electronics and Telecommunication Engineering, KIT College of Engineering, Kolhapur(Maharashtra), India.
Email: dixit.manasi@kitcoek.in b

it's called a Densely Connected Convolutional Network. DenseNet comprises Connectivity, DenseBlocks, Growth Rate, and Bottleneck layers, with a total of L(L+1)/2 direct connections for 'L' layers.

III. METHODOLOGY

3.1 DenseNet-121 as a Feature Extractor

In a conventional feed-forward Convolutional Neural Network (CNN), every convolutional layer, except the initial one that takes the input, gets information from the prior convolutional layer and generates an output feature map that is then transmitted to the subsequent convolutional layer. So, for each of the 'L' layers, there exists a direct connection to the next layer.

Connection

In every layer, the feature maps are not added together, but combined and utilized as inputs. As a result, DenseNets have fewer parameters compared to a traditional CNN of the same size, enabling feature reuse by eliminating redundant feature maps. Therefore, the lth layer is provided with the feature-maps from all previous layers, x0,...,xl-1, as its input.

where [x0,x1,...,xl-1] represents the combination of feature-maps generated by the output in all the layers prior to l (0,...,l-1). The various inputs of Hl are combined into one tensor for simplifying implementation.

Blocks with high density

Using the concatenation operation is not possible if the size of feature maps varies. Yet, a crucial aspect of CNNs involves the downsizing of layers that decreases the size of feature-maps by reducing dimensions in order to achieve faster computational speeds.

DenseNets are segmented into DenseBlocks to maintain constant feature map dimensions within a block while varying the number of filters between them. The Transition Layers, located between the blocks, decrease the channel count by half compared to the current channels.

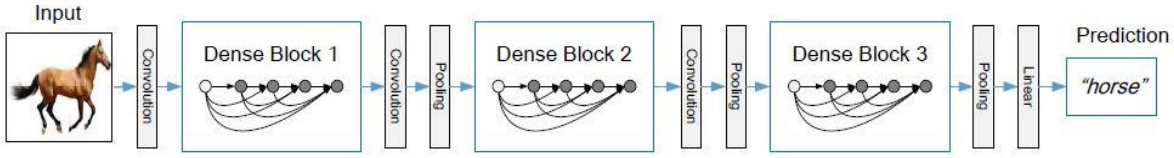

Fig.2 DenseNet Block structure

For each layer, from the equation above, H_l is defined as a composite function which applies three sequential steps: batch normalization (BN), applying a rectified linear unit (ReLU), and performing a convolution (Conv).

The image above displays a complex DenseNet model with three dense blocks. The transition layers between neighboring blocks downsample feature maps through convolution and pooling, while feature maps within the dense block remain the same size for concatenation.

Rate of growth.

The features can be considered as an overall condition of the network. The feature map expands as it goes through each dense layer, with each layer introducing 'K' additional features to the existing features. The growth rate of the network, denoted by parameter 'K', controls the quantity of information added to each layer of the network. If every function H l generates k feature maps, then the lth layer contains input feature maps are comprised of k0 channels in the input layer. Unlike current network structures, DenseNets can possess extremely slim layers.

Layers that slow down or limit the flow of a process.

Even though each layer generates k output feature-maps, the amount of inputs can be significantly large, especially for subsequent layers. Therefore, a bottleneck layer of 1x1 convolution can be added before each 3x3 convolution layer to enhance computational efficiency and speed.

Hence, DenseNet-121 is comprised of the subsequent layers:

- One convolution of size 7x7
- Fifty-eight 3x3 convolutions. 61 one by one Convolution
- 4 Average Pooling operations
- 1 Dense Layer

In brief, DenseNet-121 consists of 120 Convolution layers and 4 Average Pooling layers.

DenseNet-121 introduces dense connections between layers, enabling each layer to directly access the feature maps of all previous layers within the same dense block and transition layers. This allows deeper layers to utilize features extracted early on, while the second and third dense block layers assign fewer weights to the redundant features outputted by transition layers. Despite the final layers using the weights of the entire dense block, there may still be more high-level features further into the model, as seen in experiments. This high level of connectivity results in enhanced flow of gradients and potentially more complex feature representations. We employed a pre-

trained DenseNet-121 model for our image captioning system, where we excluded the ultimate fully connected layer and gathered features from the final convolutional block. These characteristics act as the language generation network's input.

3.2 Model for Generating Captions

The suggested model for generating image captions utilizes an encoder-decoder structure.

Encoder: DenseNet-121 generates visual characteristics, forming a feature vector as the first input for the decoder.

Decoder: An LSTM, which is a type of recurrent network, analyzes the feature vector in order to produce words one by one, ultimately creating a sentence that describes the image. The LSTM uses the DenseNet features to produce words in a step-by-step manner until a specified condition is reached, such as generating a token indicating the end of a sentence.

3.3 Training and Optimization

The model underwent training using the MS COCO dataset, consisting of more than 82,000 training images, each accompanied by five reference captions. Cross-entropy loss function was utilized during training, along with an optional reinforcement learning element to enhance consistency between generated and reference captions. The Adam optimizer was used for optimization with a starting learning rate of 0.001.

Dataset and Assessment Criteria for Evaluation.

The work utilized the Flickr 8k Dataset, which includes 8,000 images, each accompanied by five captions that offer detailed descriptions of the important subjects and activities.

- A child in a pink dress is climbing up a set of stairs in an entry way.
- A girl going into a wooden building.
- A little girl climbing into a wooden playhouse A little girl climbing the stairs to her playhouse.
- A little girl in a pink dress going into a wooden cabin.

Fig.3 Sample image with caption from Flicker dataset

IV. Results and Analysis

Our results show that the DenseNet-121-based image captioning model outperformed VGG-16 and was competitive with ResNet-50, particularly in CIDEr scores, which prioritize caption uniqueness. The dense connectivity in DenseNet-121 appears to enhance feature representation, contributing to captions that more accurately describe complex images. While computationally heavier than ResNet-50, DenseNet-121 offered a favorable balance of quality and efficiency.

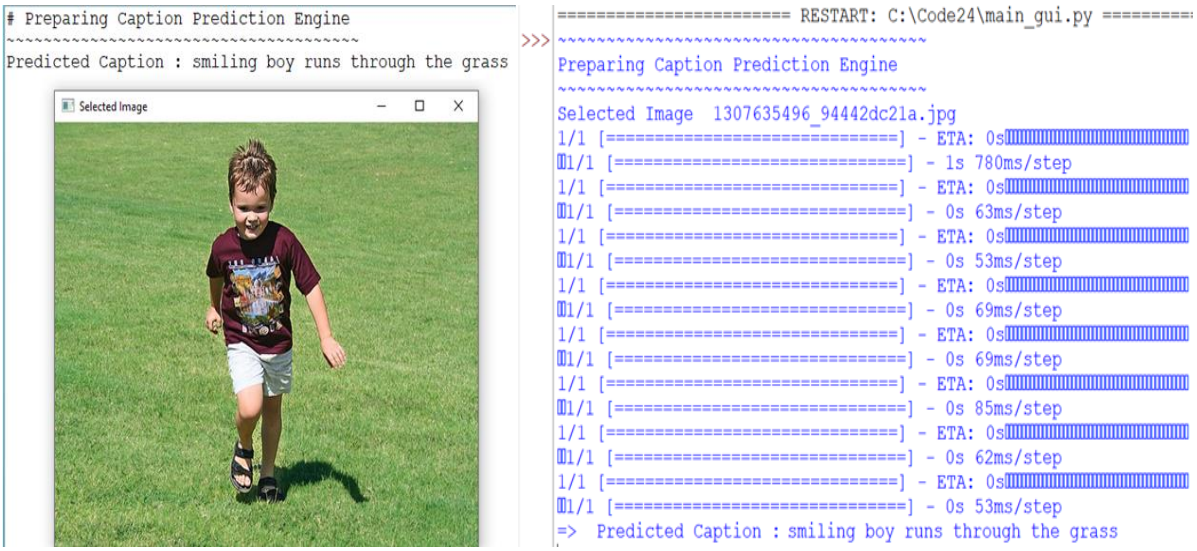

Fig.4 Output with predicted caption

BLEU Score is an evaluation metric for Machine Translation tasks. It is calculated by comparing the n-grams of machine-translated sentences to the n-gram of human-translated sentences.

V. Discussion

DenseNet-121's dense connectivity pattern encourages the reuse of learned features, which may be particularly advantageous for image captioning tasks where contextual information is critical. DenseNet-121 performed particularly well with captions for images containing intricate details or complex scenes, possibly due to its superior feature representation capabilities. However, DenseNet-121's deeper architecture increases training and inference time, which could be a limitation in real-time applications.

VI. Conclusion and Future Work

This study demonstrated the potential of DenseNet-121 as a feature extractor for image captioning tasks, achieving competitive results compared to traditional architectures. DenseNet-121's dense connections offer a more nuanced understanding of visual features, which positively impacts the quality of generated captions. Future work could explore hybrid architectures that combine DenseNet-121 with transformer-based language models, aiming to further improve the balance between computational efficiency and caption quality.

References

[1] Dr. Yashwant Dongare, Dr. Bhalchandra M. Hardas et. Al , "Deep Neural Networks for Automated Image Captioning to Improve Accessibility for Visually Impaired Users", International Journal of intelligent systems and applications in engineering – 2023.

[2] Mehmet Ali Can Ertuğrul; Sevinç İlhan Omurca , Generating Image Captions Using Deep Neural Networks , 2023 8th International Conference on Computer Science and Engineering (UBMK)

[3] Harshit Rampal , Aman Mohanty , "Efficient CNN-LSTM based Image Captioning using Neural Network Compression", arXiv:2012.09708v1 [cs.CV] 17 Dec 2020

[4] Fucheng You and Yangze Zhao , Attention Image Caption with DenseNet , IOP Conf. Series: Journal of Physics: Conf. Series 1302 (2019).

[5] Xu, K., Ba, J., Kiros, R., et al. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning (ICML), 2048–2057.

[6] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[7] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. CVPR, pp. 3156–3164.