**NANYANG TECHNOLOGICAL UNIVERSITY   
**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**Assignment for SC4001**

**AY2023-2024**

**Group ID:**

**Group members:**

|  |  |
| --- | --- |
| Name | Matric No. |
|  |  |
|  |  |
|  |  |

**1 Background Motivation**

In the rapidly evolving realm of computer vision, Convolutional Neural Networks (CNNs) have emerged as a gold standard for image recognition tasks, demonstrating unparalleled proficiency in capturing hierarchical patterns in visual data. Particularly, DenseNet201, with its densely connected architecture, has exhibited superior performance in various benchmark datasets by alleviating the vanishing gradient problem and bolstering feature propagation. However, when it comes to tasks where temporal or sequential information is pivotal, such as classifying video frames or sequences of images, CNNs alone might not be sufficient. Long Short-Term Memory networks (LSTMs) excel in these scenarios, given their unique capability to remember long-term dependencies in sequential data. Merging the prowess of DenseNet201 in feature extraction with the sequence modeling strength of LSTMs offers a promising avenue for image sequence classification. Our endeavor in this project was spurred by this synergy, aiming to harness the combined strengths of both architectures to achieve superior image classification performance, especially in scenarios where temporal or order-based information plays a critical role.

**2 Tasks**

**3 Representation**

**4 Datasets**

In this section, we take care of what we used to train and evaluate our image-captioning model’s architecture. We specifically used the *Flickr Image dataset* from (<https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset/data>), last accessed on 1 Nov 2023.

This dataset includes a folder with over 30 thousand random images. Each image has its own assigned caption that is stored in a *csv* file.

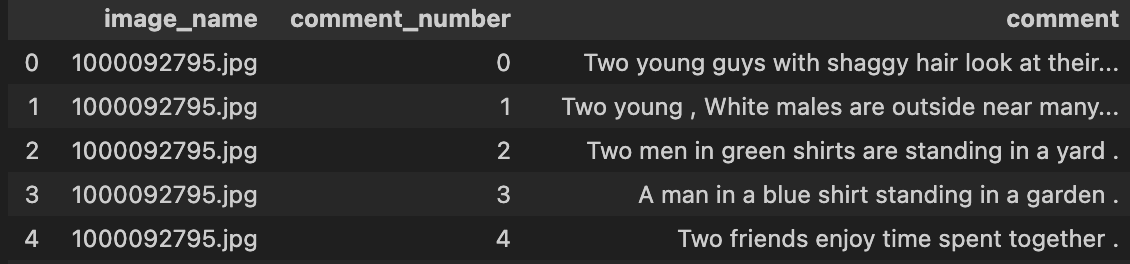


Fig 1. Example of csv

A couple of men standing outside a door

Description automatically generated

Fig 2. '1000092795.jpg'

For example, image ‘1000092795.jpg’ has 5 captions (in the ‘comment’ column):

1. Two young guys with shaggy hair look at their hands while hanging out in the yard.
2. Two young, White males are outside near many bushes.
3. Two men in green shirts are standing in a yard.
4. A man in a blue shirt standing in a garden.
5. Two friends enjoy time spent together.

**5.1 Convolution Neural Network (CNN)**

We selected 4 pretrained CNN models to test out: VGG16, DenseNet201, InceptionV3 and ResNet50. These models will be used for feature extraction for our image dataset. As we aim to develop an image-captioning model for any image, the CNN model has to be able to recognise almost every possible object in the picture.

Our research tells us that these pretrained models are trained on datasets as large as a million images, with more than 1000 classification classes that uses very deep models with more than 200 hidden layers. It is possible that we reduce the scope of our image-captioning model to let it classify a smaller number of objects, for example a model to recognise the type of clothing. However, we figured that will reduce our project to a simple classification problem where we only need to detect a few classes, making things a lot more trivial.

Instead, we opted to focus on the analysis of using the aforementioned models.

**5.2 Long Short Term Memory (LSTM)**

Our project requires the image-captioning model to generate a logical and grammatically sound sentence that describes the input image. As LSTM models are useful in sequence prediction,

Why LSTM is used here. Pros Cons

Comparing with RNN vs CNN-LSTM

**6 Analysis**

To evaluate the CNN models, we opt to use comprehensive indicators (Yang et al., 2021) that can be used for image classification analysis. According to (Yang et al., 2021), some models performed well on recall but poorly according to the accuracy and precision.

A math equations on a white background

Description automatically generated

Fig 3. Calculation of comprehensive indicator

Hence to

**7 Conclusion**

**8 References**

Yang, Y., Zhang, L., Du, M., Bo, J., Liu, H., Ren, L., Li, X., & Deen, M. J. (2021). A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions. Computers in biology and medicine, 139, 104887. https://doi.org/10.1016/j.compbiomed.2021.104887