GROUP 02

Picture This:

Evaluating the Effectiveness of CNN + LSTM and ResNet + GRU

Group Members

RUCHI KAPADIWALA
002772936

APARNA CHAVAN
002799851

ASHI TYAGI
002706544

PARTH KALANI
002766306

BACK TO AGENDA

Agenda

Problem Statement

Understanding the Dataset

CNN Model

LSTM Model

GRU Model

ResNet Model

Technical Approach

Results

Comparision

Conclusion

GROUP 02

Problem Statement

The problem we are tackling today is the task of automated image captioning using two different approaches, namely CNN + LSTM and ResNet + GRU.

The aim is to compare the performance of these methods on the Flickr 8k Dataset, using pretrained CNN/ResNet models to demonstrate the approaches.

The comparison will be based on the BLEU scores, and the goal is to evaluate the effectiveness of these approaches in generating accurate and informative captions for images.

BACK TO AGENDA

Understanding the Dataset

	image	caption
0	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set o...
1	1000268201_693b08cb0e.jpg	A girl going into a wooden building .
2	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse .
3	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playh...
4	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a woo...

A dog is running along a beach on a sunny day .

The boys dressed in athletic wear perform exercises on the grass .

A man 's sillouhette at sunset .

A boy wears plastic toy teeth and green plastic toy glasses which are much too small for him .

A girl dressed in orange flies in the air while jumping on the bed .

GROUP 02

DATA DESCRIPTION

Flickr8 Dataset

It contains a collection of 8,000 images, each with 5 different captions written by human annotators, resulting in a total of 40,000 captions.

The dataset is split into three subsets: 6,000 images for training, 1,000 images for validation, and 1,000 images for testing.

Flickr8k dataset is a popular benchmark dataset for the task of automated image captioning.

BACK TO AGENDA

GROUP 02

DATA DESCRIPTION

Flickr8 Dataset

BACK TO AGENDA

The images in the dataset cover a wide range of categories, including people, animals, landscapes, and objects.

The dataset is useful for evaluating the performance of image captioning models, as it provides a standardized benchmark for comparison.

Data Preprocessing

Preprocessing the captions by converting them to lowercase, removing special characters and whitespace, joining the words with spaces, and adding startseq and endseq tokens to indicate the beginning and the ending of a sentence

```
# before preprocess of text
mapping['1000268201_693b08cb0e']
```

```
['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 .']
```

```
# after preprocess of text
mapping['1000268201_693b08cb0e']
```

```
['startseq child in pink dress is climbing up set of stairs in an entry way endseq',
 'startseq girl going into wooden building endseq',
 'startseq little girl climbing into wooden playhouse endseq',
 'startseq little girl climbing the stairs to her playhouse endseq',
 'startseq little girl in pink dress going into wooden cabin endseq']
```

BACK TO AGENDA

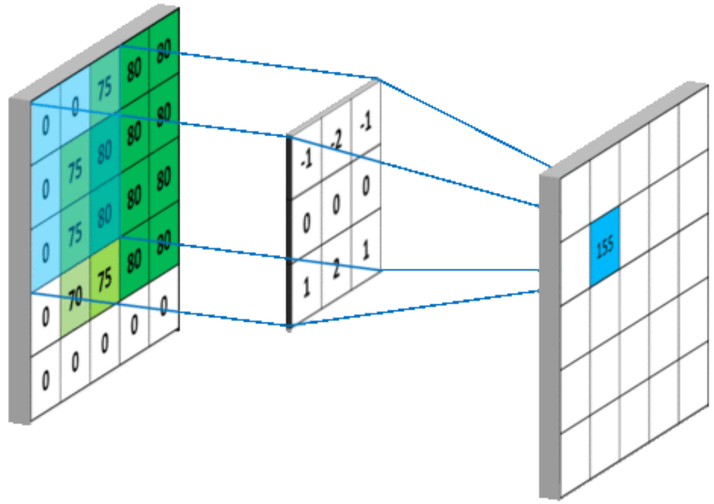
What is CNN Model?

01 CNN stands for Convolutional Neural Network, which is a type of deep learning model that is commonly used for image and video processing tasks.

02 CNN models are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

03 Convolutional layers perform a set of convolutions with a set of learnable filters, which helps to detect features such as edges, corners, and shapes in the image.

04 Some popular CNN models include AlexNet, VGGNet, GoogLeNet, ResNet, and InceptionNet.



BACK TO AGENDA

What is LSTM Model?

01

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that is commonly used for processing sequential data.

02

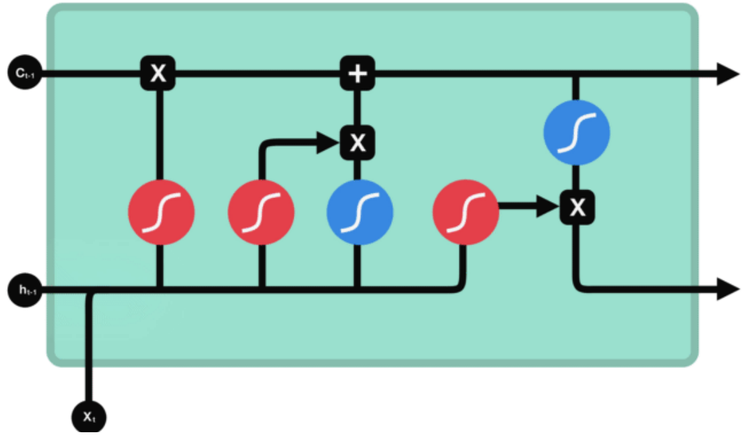
LSTM models are controlled by three gates - the input gate, the forget gate, and the output gate - that regulate the flow of information in and out of the cells.

03

The input gate determines which information to let in, the forget gate determines which information to forget, and the output gate determines which information to output.

04

This architecture allows the LSTM model to selectively remember or forget information over time, making it well-suited for processing sequential data.



BACK TO AGENDA

CNN +LSTM

01 To perform Image Captioning we will require two deep learning models combined into one for the training purpose

02 CNNs extract the features from the image of some vector size aka the vector embeddings. The size of these embeddings depend on the type of pretrained network being used for the feature extraction

03 LSTMs are used for the text generation process. The image embeddings are concatenated with the word embeddings and passed to the LSTM to generate the next word

04 For a more illustrative explanation of this architecture check the Modelling section for a picture representation

BACK TO AGENDA

High Level View of the Model

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None, 1920)]	0	[]
dense (Dense)	(None, 256)	491776	['input_2[0][0]']
input_3 (InputLayer)	[(None, 34)]	0	[]
reshape (Reshape)	(None, 1, 256)	0	['dense[0][0]']
embedding (Embedding)	(None, 34, 256)	2172160	['input_3[0][0]']
concatenate (Concatenate)	(None, 35, 256)	0	['reshape[0][0]', 'embedding[0][0]']
lstm (LSTM)	(None, 256)	525312	['concatenate[0][0]']
dropout (Dropout)	(None, 256)	0	['lstm[0][0]']
add (Add)	(None, 256)	0	['dropout[0][0]', 'dense[0][0]']
dense_1 (Dense)	(None, 128)	32896	['add[0][0]']
dropout_1 (Dropout)	(None, 128)	0	['dense_1[0][0]']
dense_2 (Dense)	(None, 8485)	1094565	['dropout_1[0][0]']

Total params: 4,316,709
Trainable params: 4,316,709
Non-trainable params: 0

BACK TO AGENDA

High Level View of the Model

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None, 1920)]	0	[]
dense (Dense)	(None, 256)	491776	['input_2[0][0]']
input_3 (InputLayer)	[(None, 34)]	0	[]
reshape (Reshape)	(None, 1, 256)	0	['dense[0][0]']
embedding (Embedding)	(None, 34, 256)	2172160	['input_3[0][0]']
concatenate (Concatenate)	(None, 35, 256)	0	['reshape[0][0]', 'embedding[0][0]']
lstm (LSTM)	(None, 256)	525312	['concatenate[0][0]']
dropout (Dropout)	(None, 256)	0	['lstm[0][0]']
add (Add)	(None, 256)	0	['dropout[0][0]', 'dense[0][0]']
dense_1 (Dense)	(None, 128)	32896	['add[0][0]']
dropout_1 (Dropout)	(None, 128)	0	['dense_1[0][0]']
dense_2 (Dense)	(None, 8485)	1094565	['dropout_1[0][0]']

Total params: 4,316,709
Trainable params: 4,316,709
Non-trainable params: 0

BACK TO AGENDA

What is ResNet?

01 ResNet is a type of deep neural network architecture designed to address the problem of vanishing gradients that can occur when training very deep neural networks.

02 The key innovation of the ResNet architecture is the use of skip connections (also called residual connections) between layers of the network.

03 Residual connections allow gradients to flow more easily through the network, which makes it easier to train very deep networks.

04 There are several variants of the ResNet architecture, including ResNet-50, ResNet-101, and ResNet-152, which differ in the number of layers and other architectural details.

BACK TO AGENDA

What is GRU?

01

GRU (Gated Recurrent Unit) is a type of recurrent neural network (RNN) architecture that was introduced in 2014 as a simpler alternative to the more complex LSTM (Long Short-Term Memory) architecture.

02

GRUs are designed to address the problem of vanishing gradients that can occur in standard RNNs when they are trained on long sequences of data.

03

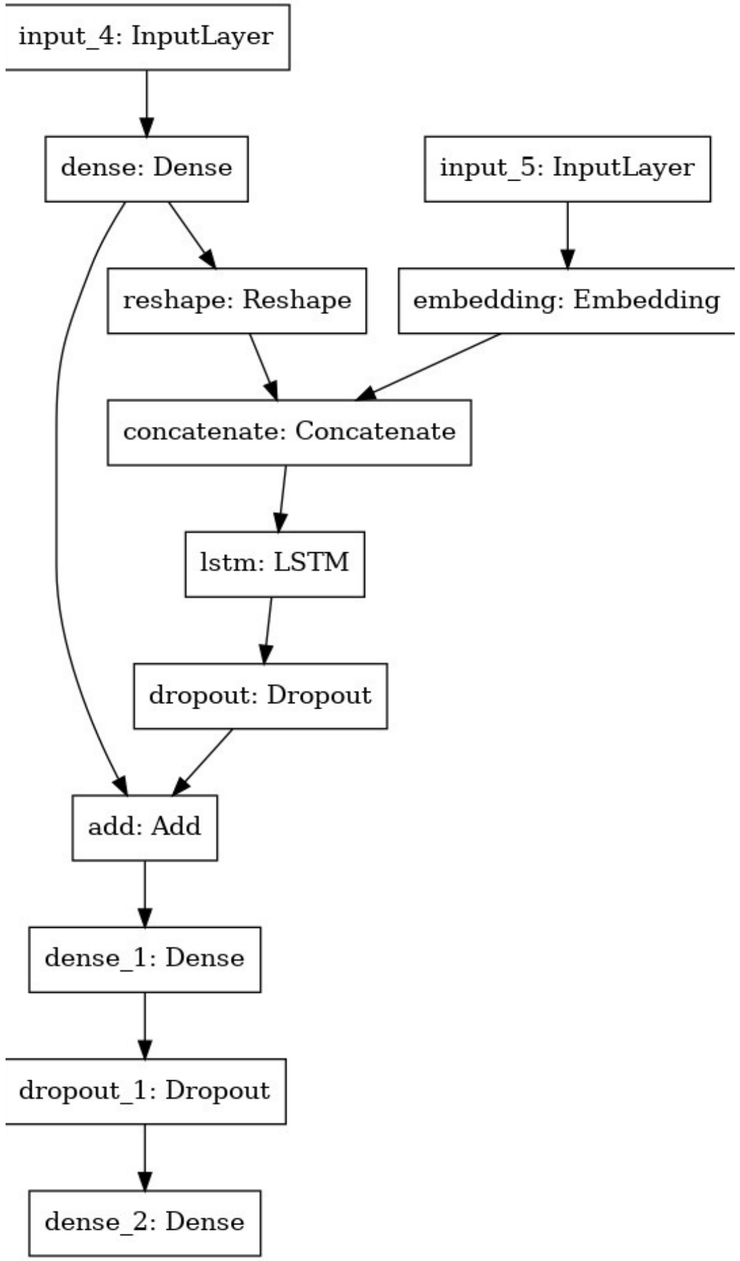
GRUs use gating mechanisms to control the flow of information through the network, with a reset gate and an update gate determining how much of the previous state to forget and how much of the new state to retain.

04

GRUs have been found to be faster to train and require fewer parameters than LSTMs, while achieving similar performance on many tasks.

BACK TO AGENDA

Architecture



BACK TO AGENDA

Technical Approach

01

Data Preprocessing:

a. Extract features from the images in the directory using the VGG16 model and preprocess the image for the VGG model.

b. Store the extracted features in a dictionary and save the dictionary as a pickle file.

c. Load the features from the pickle file and the captions from the captions.txt file.

02

Split into training and testing:

1. Split the image IDs into train and test sets.
2. Create a data generator to get data in batches and encode the sequences.

BACK TO AGENDA

Technical Approach

03

The first part of the model is a CNN that extracts features from the input images. The CNN architecture used in this code consists of three convolutional layers with 32, 64, and 128 filters respectively, each followed by a MaxPooling2D layer to downsample the output.

04

ResNet152v2 is a convolutional neural network architecture used for image recognition tasks. It has 152 layers and is an improvement over the original ResNet152 model, using additional techniques such as batch normalization and improved residual connections. ResNet152v2 has achieved state-of-the-art performance on various image recognition benchmarks.

BACK TO AGENDA

Technical Approach

05

The second part of the model is an LSTM network that takes the 1D feature vector from the CNN as input and processes it to predict the classification of the image.

06

GRU (Gated Recurrent Unit) is a type of recurrent neural network (RNN) architecture that uses gating mechanisms to improve long-term dependencies learning and prevent vanishing gradient problem. It is similar to LSTM (Long Short-Term Memory) but has fewer parameters and can be faster to train.

07

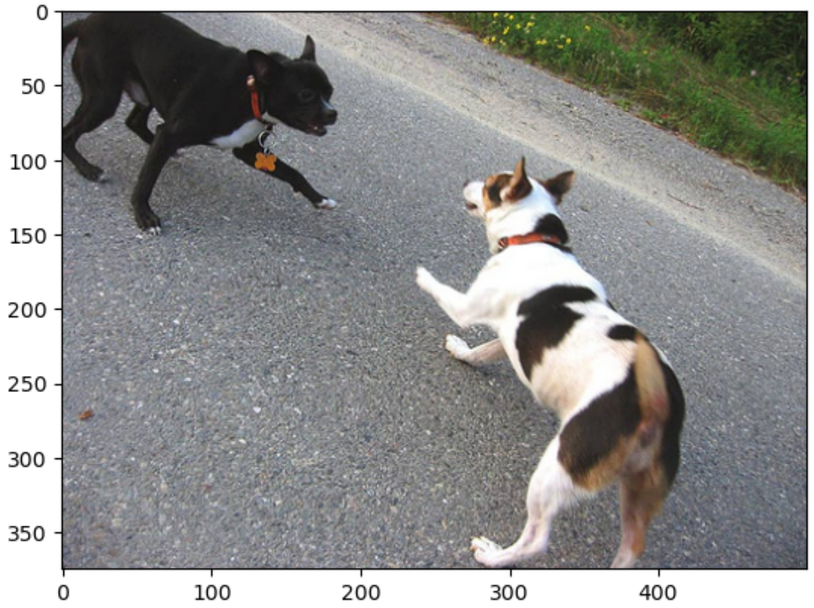
Model Testing:

a. Load the saved model weights.
b. Use the model to generate captions for new images.
c. Evaluate the generated captions using metrics such as BLEU

BACK TO AGENDA

Results - CNN + LSTM

```
---------------------Actual---------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
--------------------Predicted--------------------
startseq two dogs playing with each other on the grass endseq
```

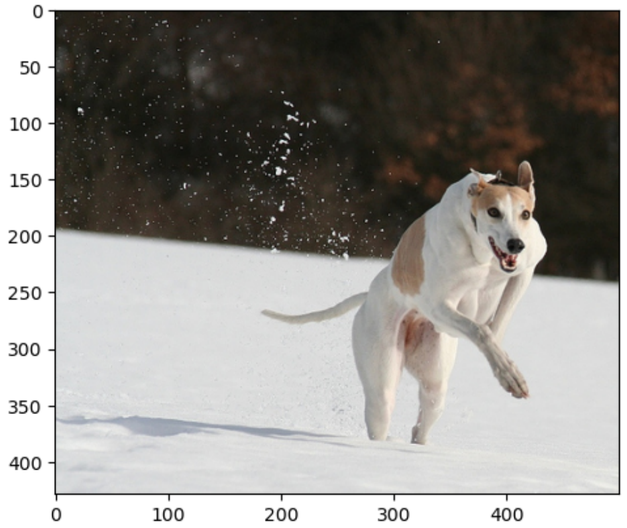



```
---------------------Actual---------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
--------------------Predicted--------------------
startseq two children are sitting in front of colorful paint and painting endseq
```

BACK TO AGENDA

Results - ResNet + GRU

```
startseq brown and white dog is running through the snow endseq
startseq dog is running in the snow endseq
startseq dog running through snow endseq
startseq white and brown dog is running through snow covered field endseq
startseq the white and brown dog is running over the surface of the snow endseq
predicted statement
startseq two dogs are running in the snow endseq
```




```
startseq crowd watching air balloons at night endseq
startseq group of hot air balloons lit up at night endseq
startseq people are watching hot air balloons in the park endseq
startseq people watching hot air balloons endseq
startseq seven large balloons are lined up at nighttime near crowd endseq
predicted statement
startseq two people are standing in front of building endseq
```

BACK TO AGENDA

BLEU Score

The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of machine translation outputs by comparing them with one or more reference translations.
The score ranges from 0 to 1, with 1 being a perfect match with the reference translations

CNN + LSTM

BLEU-1: 0.531055
BLEU-2: 0.307995

ResNet + GRU

BLEU-1: 0.3809523809523809
BLEU-2: 0.6172133998483676

BACK TO AGENDA

Conclusion

- Based on the given results, it can be concluded that the **CNN + LSTM model performs better than the ResNet + GRU model** for image caption prediction using the flickr8 dataset. The **CNN + LSTM model** achieved a higher **BLEU-1 score (0.53)** compared to the **ResNet + GRU model (0.38)**, indicating that the CNN + LSTM model generated **more accurate and relevant captions for the images**.

- However, the **ResNet + GRU model achieved a higher BLEU-2 score (0.61)** than the **CNN + LSTM model (0.30)**, which suggests that the ResNet + GRU model **performed better in generating captions with better syntactic structure and n-gram overlap.**

BACK TO AGENDA

THANK YOU!

CAPTION: POTTED PLANT ON CHAIR ON BOOKS