Image Captioning Model Performance

The performance of the different models will be shown below by displaying each image and showing what caption is generated by each of the models. All the models below use the VGG model to initially encode the image.

The different models are:-

LSTM Model : This is a simple RNN model consisting of only LSTM's. Here the image encoding obtained from our VGG model is used as the initial state of the first LSTM and the starting vector input to the LSTM is the one hot encoding of <start> token. The subsequent new words are then automatically generated by the model and used as input to the next LSTM unit. I had limited the max words in a sentence generated by the model of one image to 20 words.

Transformer Decoder Model : This model consists of a transformer decoder which decodes the next word in the sentence using the previously generated words and the image encoding output of the VGG model. The previously generated words first go through self attention to get the general idea of which words to focus on while generating the next word. The output of this self attention is used as the query input in the Multi Head Attention layer. The output of the VGG model is used as key and value input to the Multi Head Attention layer. There are also feed forward networks after self attention and Multi Head Attention layers. Finally a last fully connected layer is used to generate the token for the next word in the sentence.

Transformer Encoder and Decoder Model : This model is similar to the Transformer Decoder Model. This model has an extra image encoding transformer model which further encodes the output of the VGG model. The input to this Transformer Encoder is the sequence we get by copying the VGG models output 20(this is the hyperparameter I used here) times.

Like the transformer decoder layer of the previous model, the Image Encoding Transformer layer of this model also has a self attention layer and a feed forward network. Now the output of this encoding Transformer layer is used as the key and value input to the Transformer Decoder Layer. This then generates the next word using the words generated till now.

Transformer Model using Image Blocks and RNN : This is a model consisting of a local VGG model, a global VGG model, an LSTM and a Transformer Decoder layer(Like the first model). The image is first divided into 9(a chosen hyperparameter) 3X3 local regions. Each local block is resized to be appropriately input to the local VGG model and the output of all regions creates a sequence of vectors which is then input into the LSTM Model. There is also a separate global VGG model whose input is the entire image and whose output is concatenated with the final output of the previous LSTM. There is also a fully connected layer added after this and the output is then used as the key and value input to the Transformer decoder layer. This then generates the next word using the words generated till now.

Transformer Model using Image Blocks : This model is exactly like the previous model (Transformer Model using Image Blocks and RNN) where instead of adding an RNN layer after obtaining the sequence from the local VGG model, the sequence is directly concatenated with the result of the global VGG model. Then after adding the fully connected layer, the result is used as the key and value input in the Transformer decoder layer.

Transformer Model using Image Blocks and 1DCNN : This model is exactly like the previous model (Transformer Model using Image Blocks and RNN) where instead of adding an RNN layer after obtaining the sequence from the local VGG model, a 1DCNN layer is added which and the sequence obtained is directly concatenated with the result of the global VGG model. Then after adding the fully connected layer, the result is used as the key and value input in the Transformer decoder layer.

The Rouge1 Score for all these models

Rouge1 Score	LSTM Model	Transformer Decoder	Transformer Encoder and Decoder	Transformer using Image Blocks	Transformer using Image Blocks and RNN	Transformer using Image Blocks and 1DCNN
COCO Dataset	0.446	0.465	0.523	----	----	---
RSICD Dataset	0.601	0.605	0.607	0.591	0.608	0.564
UCM Dataset	0.658	0.706	0.681	0.679	0.673	0.596

Performance on COCO Dataset for first 3 models

First showing the captions generated for images which were in the training set for each model.

1.

LSTM Model : a group of people standing on top of a snow covered slope.
Transformer Decoder : a group of people standing on top of a snow covered slope.
Transformer Encoder and Decoder : a group of people on a snowy slope

2.

LSTM Model : a group of zebras are standing in a field
Transformer Decoder : a zebra standing in a field of glass
Transformer Encoder and Decoder : two giraffes are standing in the grass near a tree

3.

LSTM Model : a man is skiing down a snowy hill
Transformer Decoder : a man is standing on a skateboard in the snow
Transformer Encoder and Decoder : a man standing on a snow covered slope

Now showing the captions generated for images which were not in the training set

1.

LSTM Model : A living room with a couch chair and fireplace
Transformer Decoder : A living room with a couch and a table
Transformer Encoder and Decoder : A living room with a couch and a television

2.

LSTM Model : A man is playing tennis on a tennis court
Transformer Decoder : A man is playing tennis on a tennis court
Transformer Encoder and Decoder : A man is playing tennis on a court

3.

LSTM Model : A group of people standing around a table with food
Transformer Decoder : A group of people standing around a table with a cake
Transformer Encoder and Decoder : A group of people standing around a table with a large cake

4.

LSTM Model : A man holding a hot dog in his left hand
Transformer Decoder : A man in a suit and tie holding a cell phone to his ear
Transformer Encoder and Decoder : A woman holding a hot dog in her hand

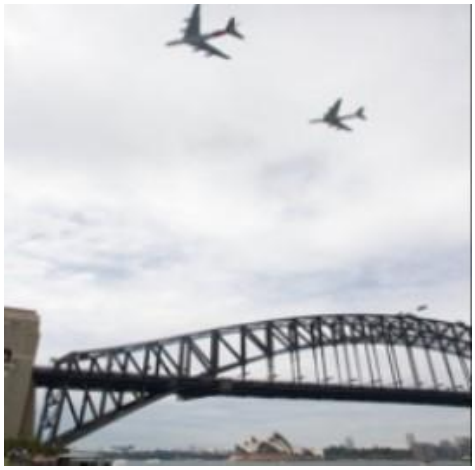
5.

LSTM Model : A train is on the tracks in a rural area
Transformer Decoder : A train travelling down train tracks next to a building
Transformer Encoder and Decoder : A street sign that is on a pole

6.

LSTM Model : A bus is parked on the side of a road
Transformer Decoder : A double decker bus is parked on the street
Transformer Encoder and Decoder : A red double decker bus is travelling down the street.

7.



LSTM Model : A large airplane is parked on the runway
Transformer Decoder : A large airplane flying over a large body of water
Transformer Encoder and Decoder : A plane flying over a body of water

Performance on RSICD Dataset for all models

First showing the captions generated for images which were in the training set for each model.

1.

LSTM Model : many buildings and green trees are in a playground with a railway station
Transformer Decoder : many green trees and some green trees are in two sides of a railway station
Transformer Encoder and Decoder : many buildings and some green trees are in two sides of a railway station
Transformer Model with Blocks and RNN : many buildings are in two sides of a railway station
Transformer Model using Image Blocks : many buildings and some green trees are in two sides of a railway station
Transformer Model using Image Blocks and 1DCNN : many buildings and some green trees are in two sides of a railway station

2.

LSTM Model : many buildings and some green trees are in an industrial area
Transformer Decoder : many buildings and some green trees are in an industrial area
Transformer Encoder and Decoder : many buildings are in an industrial area
Transformer Model with Blocks and RNN : many buildings are in an industrial area

Transformer Model using Image Blocks : many buildings and some green trees are in an industrial area
Transformer Model using Image Blocks and 1DCNN : many buildings and some green trees are in an industrial area

3.

LSTM Model : some storage areas are near a large lot
Transformer Decoder : many cars are parked in a parking lot near a road
Transformer Encoder and Decoder : many storage tanks are near a river
Transformer Model with Blocks and RNN : many storage tanks are in a factory
Transformer Model using Image Blocks : many storage tanks are near a river
Transformer Model using Image Blocks and 1DCNN : many storage tanks are near a river

4.

LSTM Model : a playground is near a road with some green trees
Transformer Decoder : a playground is surrounded by some green trees and many buildings
Transformer Encoder and Decoder : a playground is near several buildings and green trees
Transformer Model with Blocks and RNN : a football field is near several green trees and a building
Transformer Model using Image Blocks : a playground is surrounded by some green trees and buildings
Transformer Model using Image Blocks and 1DCNN : a playground is surrounded by some green trees and buildings

5.

LSTM Model : many green trees and some buildings are in a park with a pond
Transformer Decoder : many green trees and some buildings are in a park
Transformer Encoder and Decoder : many green trees and some buildings are in a park
Transformer Model with Blocks and RNN : many buildings and green trees are around a pond in a park
Transformer Model using Image Blocks : many green trees and some buildings are in a park
Transformer Model using Image Blocks and 1DCNN : many green trees and some buildings are in a park

6.

LSTM Model : a large building is near a road
Transformer Decoder : a playground is surrounded by some green trees and many buildings
Transformer Encoder and Decoder : a large building is near a river
Transformer Model with Blocks and RNN : a white center building is near a river
Transformer Model using Image Blocks : a large number of trees were planted around the stadium
Transformer Model using Image Blocks and 1DCNN : a large number of trees were planted around the house

Now showing the captions generated for images which were not in the training set

1.

LSTM Model : It is a piece of green meadow
Transformer Decoder : It is a piece of yellow desert
Transformer Encoder and Decoder : It is a piece of yellow desert
Transformer Model with Blocks and RNN : It is a piece of green meadow
Transformer Model using Image Blocks : it is a piece of yellow desert
Transformer Model using Image Blocks and 1DCNN : it is a piece of yellow desert

2.

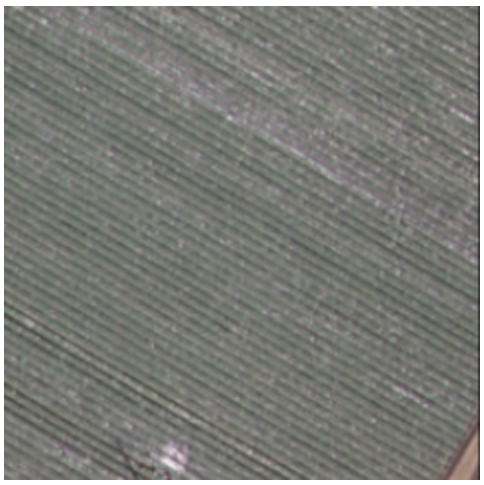
LSTM Model : many cars are parked in a parking lot

Transformer Decoder : many cars are parked in a parking lot
Transformer Encoder and Decoder : many cars are parked in a parking lot
Transformer Model with Blocks and RNN : many cars are parked in a parking lot near a building and several green trees
Transformer Model using Image Blocks : many cars are parked in a parking lot near several green trees
Transformer Model using Image Blocks and 1DCNN : many cars are parked in a parking lot

Performance on UCM Dataset for all models

First showing the captions generated for images which were in the training set for each model.

1.



LSTM Model : there is a piece of farmland
Transformer Decoder : It is a piece of farmland
Transformer Encoder and Decoder : Its is a piece of farmland
Transformer Model with Blocks and RNN : It is a piece of farmland
Transformer Model using Image Blocks : it is a piece of farmland
Transformer Model using Image Blocks and 1DCNN : there is a piece of farmland

2.

LSTM Model : an airplane is is with the the
Transformer Decoder : an airplane is surrounded with some cars in the airport
Transformer Encoder and Decoder : an airplane is surrounded with some cars in the airport
Transformer Model with Blocks and RNN : an airplane is stopped at the airport
Transformer Model using Image Blocks : an airplane is stopped at the airport
Transformer Model using Image Blocks and 1DCNN : there are some buildings with white roofs

3.

LSTM Model : there are some buildings with roofs roofs
Transformer Decoder : there are some buildings with orange roofs
Transformer Encoder and Decoder : there are some buildings with grey roofs
Transformer Model with Blocks and RNN : there are some buildings with grey roofs
Transformer Model using Image Blocks : there are some buildings with grey roofs
Transformer Model using Image Blocks and 1DCNN : many mobile homes arranged in lines in the mobile home park

4.

LSTM Model : lots of plants scattered in the loess ground
Transformer Decoder : lots of plants scattered in the loess ground
Transformer Encoder and Decoder : lots of plants scattered in the loess ground
Transformer Model with Blocks and RNN : lots of plants scattered on the ground
Transformer Model using Image Blocks : there are some grey plants scattered on the ground
Transformer Model using Image Blocks and 1DCNN : there are some grey plants scattered on the ground

5.

LSTM Model : a medium residential area with houses houses and goes through through

Transformer Decoder : there are lots of houses with grey and black roofs arranged neatly

Transformer Encoder and Decoder : a medium residential area with houses and trees

Transformer Model with Blocks and RNN : there are two tennis courts arranged neatly and surrounded by some plants

Transformer Model using Image Blocks : there are some buildings with grey roofs

Transformer Model using Image Blocks and 1DCNN : there are some buildings with grey roofs

Now showing the captions generated for images which were not in the training set

1.

LSTM Model : this is a dense with with green plants green plants
Transformer Decoder : this is a dense forest with lots of dark green trees
Transformer Encoder and Decoder : lots of dark green trees constitute a dense forest
Transformer Model with Blocks and RNN : this is a dense forest with dark green plants
Transformer Model using Image Blocks : this is a dense forest with lots of dark green trees
Transformer Model using Image Blocks and 1DCNN : a river with a dense forest on both banks of the river

2.

LSTM Model : lots of boats docked neatly at the harbor
Transformer Decoder : many boats docked neatly at the harbor and some positions are free
Transformer Encoder and Decoder : many boats docked neatly at the harbor and some positions are free
Transformer Model with Blocks and RNN : many boats docked neatly at the harbor and the water is deep blue
Transformer Model using Image Blocks : lots of boats docked neatly at the harbor and some positions are free
Transformer Model using Image Blocks and 1DCNN : lots of boats docked in lines at the harbor