Visual Recognition

Mini Project

Visual Recognition Part 2

Members:

IMT2019001—Abhinav Kamath

IMT2019012—Archit Sangal

IMT2019514—Phani Sriram

Team ID—29

Introduction

For the Mini-Project we design a CNN-LSTM system using pytroch and Keras. It can perform image captioning (A model that can take an image and produce a sentence that describes it.). We tried different CNN and LSTM architectures but this report includes the best architectures according to us. The BLEU score on the test set of Flickr8K data is taken as the judging parameter.

Flickr8K data was taken from - https://drive.google.com/drive/folders/1RQ5qHm0aVFqWDG9VBiSnXlNPl5T15Wf_?usp=sharing

Generation and Saving of Image Feature

- For generating the features from images, we use pre-trained deep CNNs. We used AlexNet provided by PyTorch.
- We generated the features using the method described above and saved them in as a pkl file.

Pre Processing of Caption

- We can notice that every image is mapped with 5 captions. We use dictionary data structure for a more efficient way of accessing the data.
- It is like a mapping for the image to the 5 captions. Image is represented by the filename and captions are stored in form a list of Strings.

Tokenization and Preprocessing

- Some features of the words of caption sentences are redundant.
- So we change every letter to a lower case, remove numeric data and punctuation.
- We used 2 delimiters to signify the start and end of the sentence.
- This helps the model to find the difference of when to start and stop predicting. Then we used Word2Vec for changing the words to vectors.
- To represent a sentence we may take the average of word vectors.

Padding

- Neural networks need to have input of a fixed size when we have a batch by batch approach.
- Since the length of the captions may be of different lengths, we add zeros so that all inputs have the same shape.

Model definition

Recurrent neural networks

Fully connected networks, we assume that the examples given to us are independent of each other w.r.t each other. But here we are working with sentences hence using dense layers is not practical. Hence, we use RNN here.

An RNN is perfect as it has short memory so it can retain some context of the words. For each combination of (input layer, hidden layer, output), we use the same network at different times.

Problems with RNN and Use of LSTMs

Main problems with RNNs are namely Vanishing and Exploding Gradients. For solving this problem we use Long Short Term Memory (LSTM) concepts.

Word Embeddings

- We avoid one-hot encoding, as it takes a lot of space.
- We enter the vectors which were initialized randomly. Vectors are generated and are called Word Embeddings.

GloVe

- GloVe stands for Global Vectors. This is like a dictionary, we have the word as a key and then we get the vector corresponding to that word.
- These are pretrained vectors, on large datasets.

Architecture Details

Inputs of the Model - Image of size (224, 244, 3) and tokens of the captions.

We then extract the features and reduce them to 256 dimensions (arbitrary). This was achieved by a combination of the linear layer with ReLU activation. This is passed to get the Context Vector of dimension 512 which also needs LSTM.

We have captions which we use to get Embedding and from embedding we get LSTM of dimension 256.

Loss Function: Categorical Cross Entropy Loss

Optimizer: Adam optimizer with constant learning rate.

Context Vector is to get the probability of different outputs.

Data Generator

- We have a huge dataset. We need to use data in batches otherwise the memory resource will get exhausted.
- The generator starts running from the point it stopped last time whenever it is called.
- We can use next() for the next batch (data is divided into batches). By doing this, we can ensure that the memory resources are used with a cap over them.

GloVe Embeddings

- We downloaded a text file with word and its embeddings.
- Each line is a string with the first element as the word and the remaining as its 200 dimensional embedding values.

- We do line by line extraction and convert the string embeddings to float values. This again works like a dictionary for us.
- There may be some new words in our captions which are not in the above dictionary. We take a random vector for it.

Training the model

- For every epoch, initialize the data generator and a loss counter.
- Number of iterations on the generator - length (all_captions) or the batch size.
- By doing this we go over each image and its captions in the training dataset. Perform a learning step for the batch generated during an iteration and update relevant objects.
- Save the model.

Evaluation

We have written a translation function, and another function to quantify its performance using a performance metric called BLEU score.

Translation

- We follow the same strategy for translation as we did while training. The translate function only needs the image, which we provide as its AlexNet features.

- The startseq token starts off the prediction. Each time, we predict a word, add it to our sentence, tokenize the new sentence (and pad it) and continue the procedure.
- The loop is broken either when endseq is encountered or maximum length is reached.

BLEU scores

To compare model performance for seq2seq tasks (ones where it generates a sequence and we have to match it with reference sequences to see how well it has done), we use BLEU scores.

This function checks the correctness of the generated sentence at multiple levels, as specified by the user.

To compute BLEU scores, we will use the corpus_bleu function from nltk.translate.bleu_score. It needs 3 parameters:

- List of references: List of documents (lists as well), each document being the set of possible correct translations.
- List of hypotheses: List of predictions.
- Weights: These determine the value of K in BLEU-K scores.

The model performs surprisingly well on the training dataset, and decent on the test dataset. There could be multiple reasons to this.

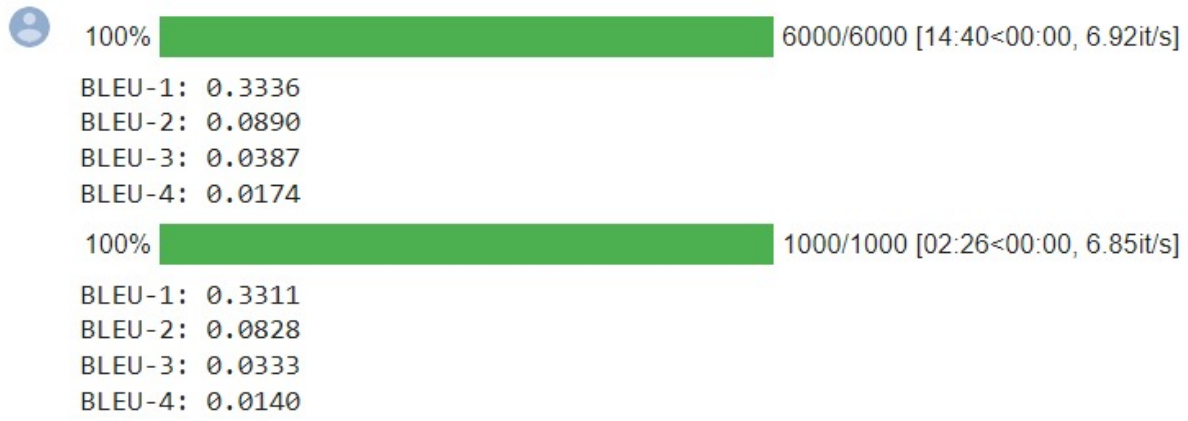
- Overfitting. There might have been too many parameters to train for the amount of data we have. Given the gap in training and test metric values

(and the brilliant performance on training data), this is the most likely case.

- There were elements in the images of test dataset that weren't encountered by the model very often in the train dataset. It didn't learn how to use the words describing them very well, and it couldn't reproduce them during testing.
- The model needs to train more.

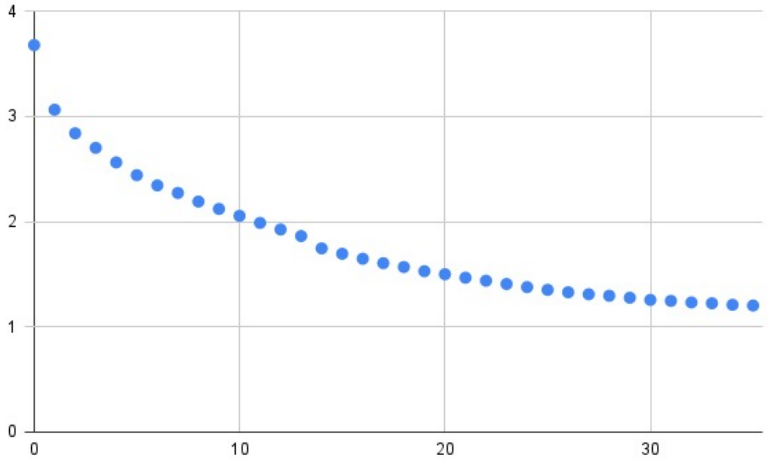
Examples and Values Obtained

Objective analysis



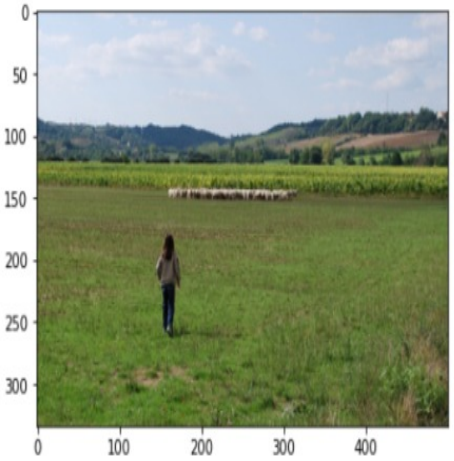
The above BLEU score is obtained for the training and the test data,

Loss Plot during training for each epoch,



Some examples of the Captioning-

Image 1



Predicted CAPTION: a woman in a blue jumpsuit popping a backpack in a field
Real CAPTION: ['begin a female walking through grass end', 'begin a girl going towards a heard of sheep end', 'begin a girl walks toward a herd of sheep in the distance end', 'begin a girl walks towards some sheep in a grassy valley end', 'begin a woman walks to the cows end']

Image 2

```
Predicted CAPTION: a dog jumps hurdle with a tennis ball in its mouth
Real CAPTION: ['begin a dog jumps up towards a woman in a car while another dog is outside the car end', 'begin an orange dog and a little white dog
with black spots are next to a white car end', 'begin a small black and white dog jumps at a woman in a white jeep while a golden dog follows beside
them end', 'begin a small dog jumps by a car to reach a lady face while another dog watches end', 'begin white and black dog jumps at someone throug
h car window end']
```

- Predicted CAPTION: a dog jumps hurdle with a tennis ball in its mouth
- Real CAPTION: ['begin a dog jumps up towards a woman in a car while another dog is outside the car end', 'begin an orange dog and a little white dog with black spots are next to a white car end', 'begin a small black and white dog jumps at a woman in a white jeep while a golden dog follows beside them end', 'begin a small dog jumps by a car to reach a lady face while another dog watches end', 'begin white and black dog jumps at someone through car window end']

Image 3

```
Predicted CAPTION: a boy plays in a pink swimsuit
Real CAPTION: ['begin a boy in a red suit plays in the water end', 'begin a boy in a red swimsuit jumps into the water to join two people end', 'beg
in a boy takes a flying leap into the water end', 'begin the boy in the red shorts jumps into the water to join other people end', 'begin the boy we
aring red shorts is jumping into the river as other children swim end']
```

- Predicted CAPTION: a boy plays in a pink swimsuit

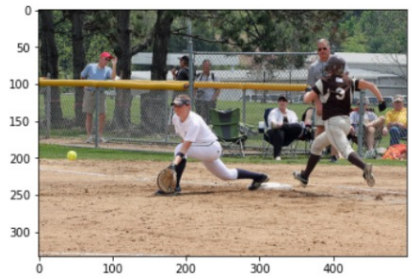
- Real CAPTION: ['begin a boy in a red suit plays in the water end', 'begin a boy in a red swimsuit jumps into the water to join two people end', 'begin a boy takes a flying leap into the water end', 'begin the boy in the red shorts jumps into the water to join other people end', 'begin the boy wearing red shorts is jumping into the river as other children swim end']

Image 4

```
Predicted CAPTION: a brown dog biting a horse leg
Real CAPTION: ['begin a little tan dog with large ears running through the grass end', 'begin a playful dog is running through the grass end', 'begi
n a small dogs ears stick up as it runs in the grass end', 'begin the small dog is running across the lawn end', 'begin this is a small beige dog ru
nning through a grassy field end']
```

- Predicted CAPTION: a brown dog biting a horse leg
- Real CAPTION: ['begin a little tan dog with large ears running through the grass end', 'begin a playful dog is running through the grass end', 'begin a small dogs ears stick up as it runs in the grass end', 'begin the small dog is running across the lawn end', 'begin this is a small beige dog running through a grassy field end']

Image 5



```
Predicted CAPTION: a group of people playing soccer on a field of field
Real CAPTION: ['begin a baseball player attempts to catch a ball while another runs towards the base end', 'begin a baseball player catches the ball
while the batter reaches the base end', 'begin a baseman tries to catch a ball while a runner tries to make the base in a community game end', 'begi
n an umpire in a baseball game crouches to catch a ball while an opposing team member runs to homebase end', 'begin two children are playing basebal
l end']
```

- Predicted CAPTION: a group of people playing soccer on a field of field
- Real CAPTION: ['begin a baseball player attempts to catch a ball while another runs towards the base end', 'begin a baseball player catches the ball while the batter reaches the base end', 'begin a baseman tries to catch a ball while a runner tries to make the

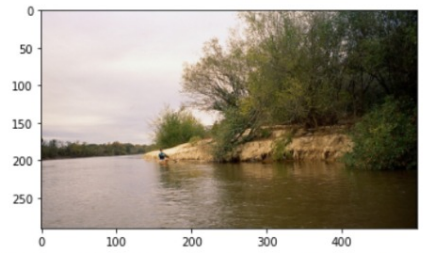
base in a community game end', 'begin an umpire in a baseball game crouches to catch a ball while an opposing team member runs to homebase end', 'begin two children are playing baseball end']

Image 6


```
Predicted CAPTION: a football player is in action among other players
Real CAPTION: ['begin a coach speaking to a football player while everyone watches end', 'begin a football player from ou talks to the coach about t
he game end', 'begin a football player is talking to his coach end', 'begin the coaching official and football player are talking in front a crowd a
nd other team members end', 'begin the coach talks it over with his quarterback during a timeout end']
```

- Predicted CAPTION: a football player is in action among other players
- Real CAPTION: ['begin a coach speaking to a football player while everyone watches end', 'begin a football player from ou talks to the coach about the game end', 'begin a football player is talking to his coach end', 'begin the coaching official and football player are talking in front a crowd and other team members end', 'begin the coach talks it over with his quarterback during a timeout end']

Image 7



```
Predicted CAPTION: two people row in lawn in the middle of the water
Real CAPTION: ['begin a man boating along a river near the shore end', 'begin a man paddling a kayak along the shore of a river end', 'begin a perso
n is riding in a canoe on a lake next to green trees end', 'begin a person is wading through a river near an eroded embankment end', 'begin a single
person rowing a small boat on a lake end']
```

- Predicted CAPTION: two people row in lawn in the middle of the water
- Real CAPTION: ['begin a man boating along a river near the shore end', 'begin a man paddling a kayak along the shore of a river end', 'begin a person is riding in a canoe on a lake next to green trees end', 'begin a person is wading through a river near an eroded embankment end', 'begin a single person rowing a small boat on a lake end']

Part—2

Introduction

Language Bias

The LSTM-based image captioning can ‘blindly’ learn the structure of the language and predict meaningful sentences, even without learning much insight into the content of the image. This is termed the “language bias” of the system.

Finding Nouns from the sentences

```
sentences = nltk.sent_tokenize(lines) #tokenize sentences
nouns = [] #empty to array to hold all nouns

for sentence in sentences:
     for word,pos in nltk.pos_tag(nltk.word_tokenize(str(sentence))):
         if (pos == 'NN' or pos == 'NNP' or pos == 'NNS' or pos == 'NNPS'):
             nouns.append(word)
```

We run the above code for the 5 captions given to us and the predicted caption.

- Using this, we get a list of Strings, which are lists of nouns.
- We can convert this to sets, say S1 and S2, and take the intersection of the sets.

Finding Verbs from the sentences

Similarly, we can get a list of Strings, which are lists of verbs. We can convert this to sets, say S3 and S4, and take the intersection of the sets.

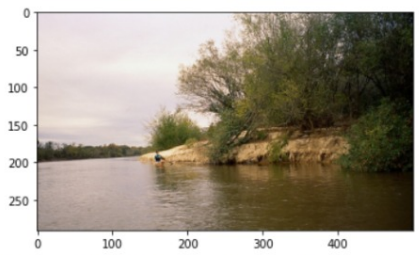
Evaluation Parameter

For the 2 intersection sets, our model's architecture predicts better, if the overlap which resulted in the intersection sets is as high as possible. i.e., the overlap between S1 and S2, and S3 and S4.

Objective and Subjective Analysis

These are some of the results that we were able to obtain after using the above analysis scheme to enquire about the bias.

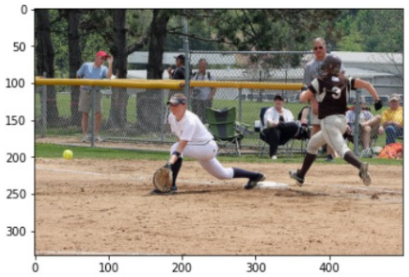
- **Image 1**



```
Predicted CAPTION: two people row in lawn in the middle of the water
Real CAPTION: ['begin a man boating along a river near the shore end', 'begin a man paddling a kayak along the shore of a river end', 'begin a perso
n is riding in a canoe on a lake next to green trees end', 'begin a person is wading through a river near an eroded embankment end', 'begin a single
person rowing a small boat on a lake end']
```

  - Nouns in the predictions : ['people', 'lawn', 'middle', 'water']
  - Nouns in the true captions : {'person', 'trees', 'kayak', 'shore', 'man', 'canoe', 'boat', 'end', 'river'}
  - Noun intersections : {'person'}

- **Image 2**



```
Predicted CAPTION: a group of people playing soccer on a field of field
Real CAPTION: ['begin a baseball player attempts to catch a ball while another runs towards the base end', 'begin a baseball player catches the ball
while the batter reaches the base end', 'begin a baseman tries to catch a ball while a runner tries to make the base in a community game end', 'begi
n an umpire in a baseball game crouches to catch a ball while an opposing team member runs to homebase end', 'begin two children are playing basebal
l end']
```

  - Nouns in the predictions : ['group', 'people', 'soccer', 'field', 'field']

- Nouns in the true captions : {'baseball', 'children', 'player', 'batter', 'runner', 'base', 'member', 'community', 'crouches', 'ball', 'game', 'team', 'end', 'baseman', 'attempts', 'umpire'}
- Noun intersections : empty set()

- **Image 3**


```
Predicted CAPTION: a football player is in action among other players
Real CAPTION: ['begin a coach speaking to a football player while everyone watches end', 'begin a football player from ou talks to the coach about t
he game end', 'begin a football player is talking to his coach end', 'begin the coaching official and football player are talking in front a crowd a
nd other team members end', 'begin the coach talks it over with his quarterback during a timeout end']
```

- Nouns in the true captions : {'front', 'talks', 'members', 'crowd', 'quarterback', 'player', 'football', 'everyone', 'speaking', 'game', 'coach', 'watches', 'coaching', 'end', 'official', 'team'}
- Nouns in the predictions : ['football', 'player', 'action', 'players']
- Noun intersections : {'football', 'player'}

- **Image 4**

```
Predicted CAPTION: a brown dog biting a horse leg
Real CAPTION: ['begin a little tan dog with large ears running through the grass end', 'begin a playful dog is running through the grass end', 'begi
n a small dogs ears stick up as it runs in the grass end', 'begin the small dog is running across the lawn end', 'begin this is a small beige dog ru
nning through a grassy field end']
```

- Nouns in the predictions : ['dog', 'horse', 'leg']
- Nouns in the true captions : {'lawn', 'dog', 'field', 'ears', 'end', 'dogs', 'grass', 'begin', 'beige'}
- Noun intersections : {'dog'}

- **Image 5**


```
Predicted CAPTION: a boy plays in a pink swimsuit
Real CAPTION: ['begin a boy in a red suit plays in the water end', 'begin a boy in a red swimsuit jumps into the water to join two people end', 'beg
in a boy takes a flying leap into the water end', 'begin the boy in the red shorts jumps into the water to join other people end', 'begin the boy we
aring red shorts is jumping into the river as other children swim end']
```

  - Nouns in the predictions : ['boy', 'pink', 'swimsuit']
  - Nouns in the true captions : {'swimsuit', 'shorts', 'children', 'suit', 'leap', 'water', 'jumps', 'end', 'river', 'boy', 'people'}
  - Noun intersections : {'swimsuit', 'boy'}