Image Captioning the COCO Dataset Using RNNs with LSTM

Jacky Dam
University of California San Diego
San Diego, California
jdam@ucsd.edu

Dylan Loe
University of California San Diego
San Diego, California
djloe@ucsd.edu

Ali Zaidi
University of California San Diego
San Diego, California
mazaidi@ucsd.edu

Abstract

In this project, we utilized recurrent neural networks (RNN) to generate captions for a given image from the data set: COCO 2015 Image Captioning Task. From this specific data set, we sampled a small portion (1/5) of the original images each containing five representative captions to create our training, validation, and test sets. With these data sets, we trained our models using the images from the training set, and evaluated the loss, BLEU-1 scores, BLEU-4 scores on the validation and test sets. Our results show that our baseline LSTM model achieved superior performance to the vanilla RNN model. In the end, we were able to create an algorithm which generates adequate predicted captions which has resemblance to the true captions for the image. Our baseline LSTM and vanilla RNN models achieved cross-entropy losses of 1.503 and 1.552, as well as BLEU-1 and BLEU-4 scores of .656 and .068 and .639 and .065, respectively, which indicated that the LSTM added to the RNN allowed for superior performance. Our experiments with stochastic and deterministic forms of caption generation illustrated that the greedy approach worked better than the random one, with BLEU scores dropping as the degree of randomness (temperature) was increased. Our attempts at varying hyperparameters like vocabulary threshold, embedding size, and hidden size, as well as the number of epochs, produced unsatisfactory performance that seemed to be overfitting, so we used the same default hyperparameters for our model with Architecture 2. Our second architecture worked slightly better than our first, with BLEU-1 and BLEU-4 scores of .664 and .075.

1 Introduction

Recurrent neural networks have been shown to perform well on sequential data with temporal structure, such as text data or frames of a video. Using a one-to-many architecture, a recurrent neural network can receive as its input an image embedding from a pre-trained CNN encoder network, and, using an RNN/LSTM as a decoder network, generate captions for that image [1]. We train our model using a fraction of the images in the COCO (Common Objects in Context) dataset, a large-scale object detection, segmentation, and image captioning dataset, with approximately 82,000 images in the training and validation sets and 3,000 in the test set, and multiple captions per image [2]. Generated captions will be evaluated using BLEU-1 and BLEU-4 scores, a metric commonly used in rating machine translation models. We consider a baseline model with default hyperparameters utilizing LSTM, a vanilla RNN model, an improvement of our baseline model after hyperparameter tuning, and lastly, an architecture which receives as its input at each time step the image embedding as well as the word generated at the previous time step. As stated in the abstract, this last approach ended up being the most successful, and was the model that we used to showcase our model's performance below.

2 Related Works

From the articles referenced below, we learned many great tips from the tutorial which helped us create our model and Pytorch library which helped us debug certain errors. Additionally, we referenced below, a research paper which worked on this specific data set which we used to cross-reference and confirm the results we got after training the algorithm.

https://bengio.abracadoudou.com/publications/pdf/vinyals_2016_pami.pdf
https://blog.floydhub.com/a-beginners-guide-on-recurrent-neural-networks-with-pytorch
https://towardsdatascience.com/automatic-image-captioning-with-cnn-rnn-aae3cd442d83
https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html
https://discuss.pytorch.org/t/proper-way-to-combine-linear-layer-after-lstm/49634/3

3 Baseline LSTM and Vanilla RNN Models (Deterministic)

All of the models we consider use a pretrained CNN, specifically ResNet50, in order to generate useful image embeddings by replacing the final softmax layer originally used for classification with a trainable linear layer. Our models were trained using "teacher forcing," i.e., using the actual word from the captions as input to the next time step rather than the word that the network generates. During testing, the model is fed the starting index and is trained to output the next word in the caption, which when fed back into the network, will be able to generate captions for yet unseen images. Our baseline model consists of a RNN augmented with a long short term memory cell (LSTM) which endows the network with "memory" that alleviates the problem of exploding and vanishing gradients typically encountered in multilayer RNNs; a vanilla RNN model is developed alongside our baseline model in order to illustrate the effects the LSTM has on the image captioning task [3].

Predicted words can be generated by the model in two ways: deterministically or stochastically. The final softmax layer produces a probability distribution over the model's vocabulary, and in the deterministic setting, it greedily picks the word with the maximum probability at each time step; the stochastic setting involves sampling words from each time step's probability distribution, smoothed with a temperature hyperparameter, in order to create novel predictions [4].

Hyperparameters for our models include a vocabulary threshold indicating the number of times a word must appear in the dataset's captions before it is discarded, an embedding size denoting the dimensionality to which we reduce one-hot encoded word vectors created from the captions, a hidden size designating the dimensionality of the linear layer through which the word vectors travel in the decoder network, and lastly, parameters pertaining to our mini-batch stochastic gradient descent procedure used during training (number of epochs, batch size, learning rate, choice of optimizer, etc.). Lastly, our models are evaluated using cross-entropy loss and BLEU scores.

4 Vanilla RNN vs. Baseline LSTM (Deterministic)

After we changed our models which utilized LSTM to Vanilla RNN, we discovered a drop in performance after switching to Vanilla RNN. Looking at the graph in figure 1 which displays loss over epochs, it's evident that LSTM performs better as we increase the number of epochs. Additionally, the difference in performance can be seen in table 2 which shows LSTM having a much higher bleu1, bleu4 score. Our results make sense because LSTM does not suffer from the problem of exploding and vanishing gradients like the Vanilla RNN model does.

Hyperparameter	Baseline LSTM	Vanilla RNN
Vocabulary Threshold	2	2
Hidden Size	512	512
Embedding Size	300	300
Number of epochs	10	10
Batch size	64	64
Learning rate	.0005	.0005
Optimizer	Adam	Adam
Training Loss	1.3929	1.4376
Validation Loss	1.5063	1.5255

Table 1: Summary of hyperparameters and losses for baseline LSTM and vanilla RNN models. (Deterministic)

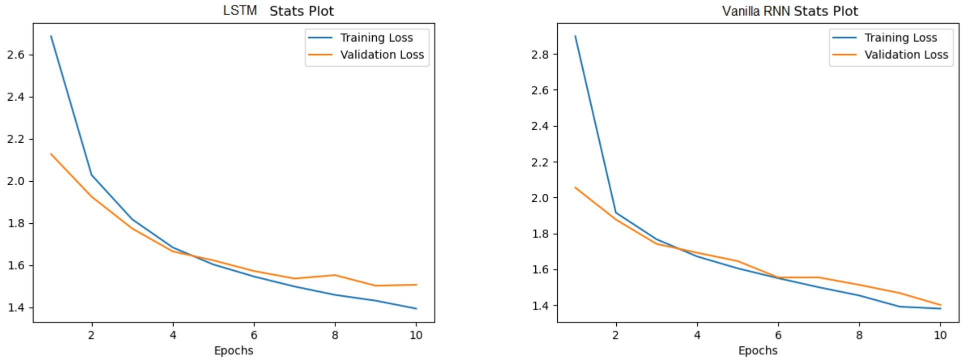


Figure 1: Training and validation accuracy and loss for baseline LSTM model and vanilla RNN model (Deterministic).

From our plots showing training and validation loss vs number of epochs, we can see that our baseline LSTM model indeed outperforms the vanilla RNN model using the same set of hyperparameters. The LSTM curve looks like it descends more smoothly and monotonically decreases, while the learning curve for the vanilla RNN looks almost piecewise linear, decreasing at practically the same rate at each epoch, and its validation curve goes up and down more than the baseline LSTM's. If we had trained both of these models for more epochs, the LSTM curve would continue to decrease (but perhaps overfit), whereas the vanilla RNN would stop learning entirely as the slope of the learning curve (a.k.a. gradient) seems to be entirely zeroed out by the 10th epoch and looks completely flat.

The cross entropy loss on the test set for the baseline LSTM model was 1.503 while the vanilla RNN achieved a loss of 1.552. The captions that the LSTM model produced were both longer and of higher quality than the vanilla RNN's.

5 Caption Generation

Model	loss	BLEU-1	BLEU-4
Baseline LSTM	1.503	0.656	0.068
Vanilla RNN	1.552	0.639	0.065

Table 2: Deterministic caption generation gives these BLEU-1 and BLEU-4 scores for our models

6 Vanilla RNN vs. Baseline LSTM (Stochastic)

In our project, the primary difference between stochastic and deterministic is the methodology of generating captions, which happens during the testing phase. With this concept in mind, we can use the same pre-trained model and simply alternate between which method of generating captions we would like to use. Generating captions deterministically takes inputs of an array of probabilities which represents the probability of the word being predicted and returns the numerical encoding of the word that has the highest probability and sends information to the embedding layer which influences the next generation of words. Generating captions stochastically works similarly to the deterministic approach and returns the numerical encoding of the word based on a certain random probability, but utilizes a parameter called temperature τ which influences the randomness.

In our experiment, we chose to vary the temperature from the choices: 0.1, 0.2, 0.7, 1.5, 2.0, and test our stochastic method on the LSTM and Vanilla RNN models. Looking at Tables 3, 4, there is a distinctive pattern which shows a decay in performance as temperature τ rises, showing us a lower bleu score at each change. In our experiment, we found that as the temperature approaches 0, the loss, bleu1, bleu4 scores start approaching the values created by the deterministic approach as shown in table 2. In a way, these results make sense because as temperature rises, the probability distribution starts becoming uniformally distributed which means that any word from the dictionary has the same probability of getting picked signifying greater randomness. For this reason, we see a significant drop in performance in both the LSTM, and Vanilla RNN models as shown in tables 3, 4 when τ = 2.

Temperature τ	loss	BLEU-1	BLEU-4
.1	1.502	0.653	0.067
.2	1.501	0.647	0.064
.7	1.498	0.563	0.039
1.5	1.514	0.182	0.010
2	1.502	0.062	0.005

Table 3: Stochastic caption generation gives these BLEU-1 and BLEU-4 scores at various temperatures for baseline LSTM model

Temperature τ	loss	BLEU-1	BLEU-4
.1	1.552	0.639	0.065
.2	1.553	0.630	0.061
.7	1.536	0.538	0.035
1.5	1.536	0.178	0.010
2	1.547	0.067	0.005

Table 4: Stochastic caption generation gives these BLEU-1 and BLEU-4 scores at various temperatures for vanilla RNN model

7 Improvement of Baseline Model

Two important hyperparameters of our model are the embedding size, which represents the length of the image feature vectors after being fed through our encoder network, ResNet50, and the hidden size, or number of hidden units in the decoder network. The embedding size comes with a bias-variance tradeoff in that it is desirable for the vectors to be small enough to allow for efficient computation, but also large enough to keep enough information for the network to produce a correct caption. The hidden size acts as an intermediate step in the sizing of the vectors, and must also have a value so that it contains useful information but is small enough to be used across many time steps. We that that increasing the vocabulary threshold from 2 to 4 would help eliminate certain rare words in the vocabulary that helped reduce its length, which allowed it to be more easily mapped to a lower dimensional space when transformed into a word embedding. Additionally, we wanted to keep the ratio of embedding size to hidden size about the same as they are in the default parameters, but also increase both of the parameters to get a larger dimensional size vector and thus more variance or information in our data. We tried an embedding size of 450 and a hidden size of 900, and running for 20 epochs instead of 10, but in the end, these changes did not allow for better caption generation and higher BLEU scores; therefore, our best model used the default parameters.

Hyperparameter	Improved LSTM
Vocabulary Threshold	4
Hidden Size	900
Embedding Size	450
Number of epochs	20
Batch size	64
Learning rate	.0005
Optimizer	Adam
Training Loss	1.22
Validation Loss	1.42

Table 5: Summary of hyperparameters and losses for improved LSTM model

Model	loss	BLEU-1	BLEU-4
Attempted Improvement of baseline	1.44	.65008	0.0789

Table 6: Deterministic caption generation gives these BLEU-1 and BLEU-4 scores for our attempted improvement

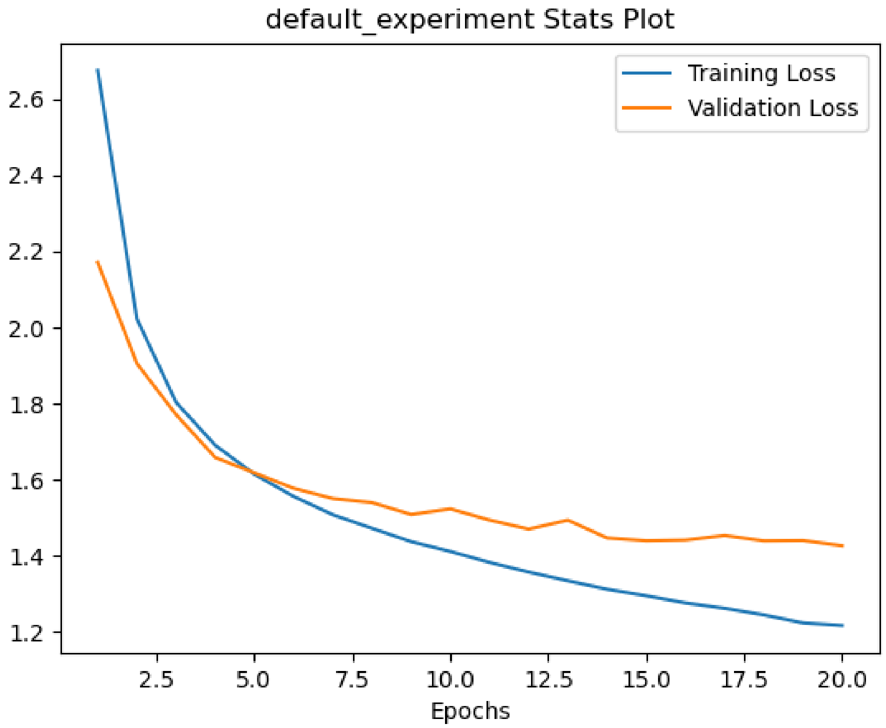


Figure 2: Training and validation loss for attempt at improvement of baseline LSTM

8 Second Model Architecture

The Second Model Architecture is very similar to First Model Architecture, except now we pass the feature vector (from the pretrained encoder) and the embedded words at each time step. We horizontally concatenate the feature vector and word embeddings into the LSTM model. For the first time step, we pass in the <pad> token for the word embedding and feature vector, so model learns to output <start>. After that, we take the output of previous time step and make it the next word embedding (next time step). We used the same hyperparameters as the improved baseline and got test loss of 1.44, BLEU-1: .663, and BLEU-4: 0.776 for 10 epochs. We didn't see major improvement compared with the improved baseline.

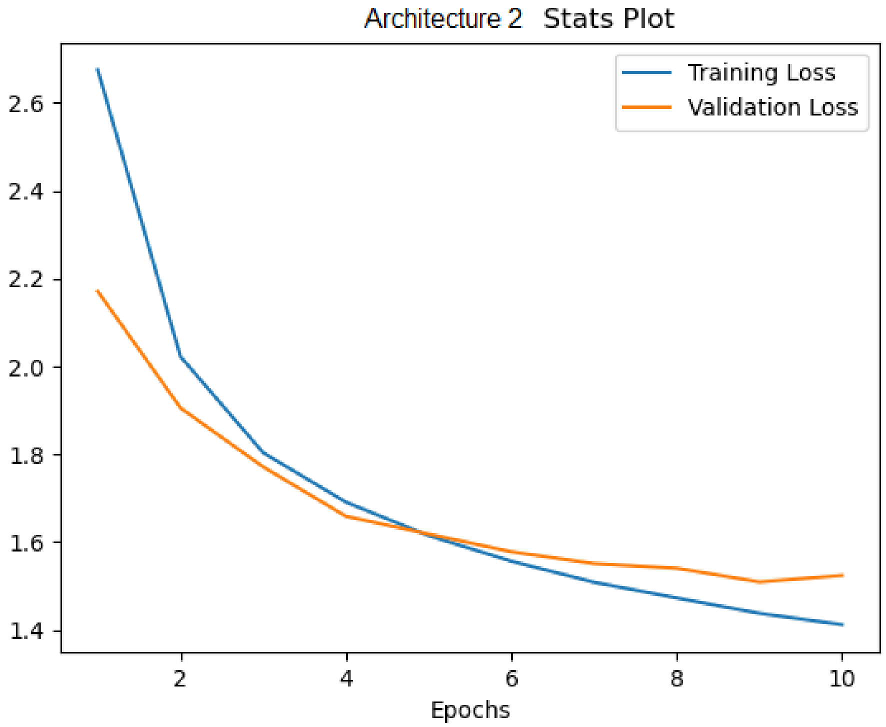


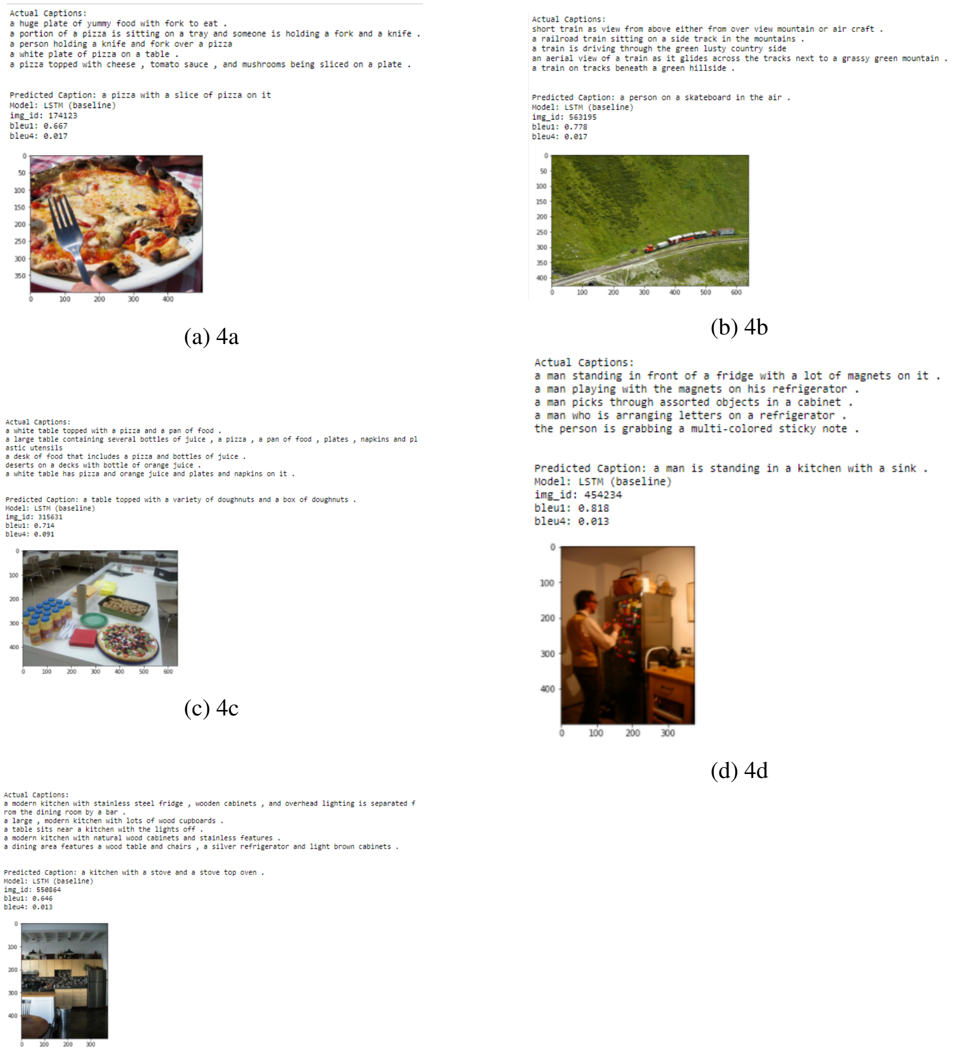
Figure 3: Training and validation loss for Architecture 2

Model	loss	BLEU-1	BLEU-4
Architecture 2	1.44	.663	0. 0776

Table 7: Deterministic caption generation gives these BLEU-1 and BLEU-4 scores for our best model, Architecture 2

9 Sample Images and Generated Captions from Best Model

Using our best model across all the experiments that we did, we can generate captions for novel images from our test set that the model has never seen. (see below images)



```
Actual Captions:
a huge plate of yummy food with fork to eat .
a portion of a pizza is sitting on a tray and someone is holding a fork and a knife .
a person holding a knife and fork over a pizza
a white plate of pizza on a table .
a pizza topped with cheese , tomato sauce , and mushrooms being sliced on a plate .


Predicted Caption: a pizza with a slice of pizza on it
Model: LSTM (baseline)
img_id: 174123
bleu1: 0.667
bleu4: 0.017
```

(a) 4a

```
Actual Captions:
short train as view from above either from over view mountain or air craft .
a railroad train sitting on a side track in the mountains .
a train is driving through the green lusty country side
an aerial view of a train as it glides across the tracks next to a grassy green mountain .
a train on tracks beneath a green hillside .


Predicted Caption: a person on a skateboard in the air .
Model: LSTM (baseline)
img_id: 563195
bleu1: 0.778
bleu4: 0.017
```

(b) 4b

```
Actual Captions:
a white table topped with a pizza and a pan of food .
a large table containing several bottles of juice , a pizza , a pan of food , plates , napkins and pl
astic utensils
a desk of food that includes a pizza and bottles of juice .
deserts on a decks with bottle of orange juice .
a white table has pizza and orange juice and plates and napkins on it .


Predicted Caption: a table topped with a variety of doughnuts and a box of doughnuts .
Model: LSTM (baseline)
img_id: 315631
bleu1: 0.714
bleu4: 0.091
```

(c) 4c

```
Actual Captions:
a man standing in front of a fridge with a lot of magnets on it .
a man playing with the magnets on his refrigerator .
a man picks through assorted objects in a cabinet .
a man who is arranging letters on a refrigerator .
the person is grabbing a multi-colored sticky note .


Predicted Caption: a man is standing in a kitchen with a sink .
Model: LSTM (baseline)
img_id: 454234
bleu1: 0.818
bleu4: 0.013
```

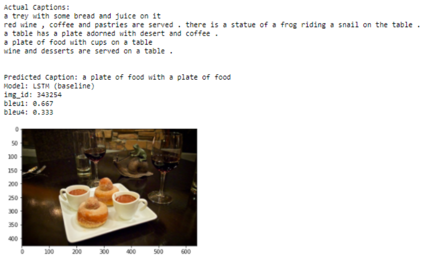
(d) 4d

```
Actual Captions:
a modern kitchen with stainless steel fridge , wooden cabinets , and overhead lighting is separated f
rom the dining room by a bar .
a large , modern kitchen with lots of wood cupboards .
a table sits near a kitchen with the lights off .
a modern kitchen with natural wood cabinets and stainless features .
a dining area features a wood table and chairs , a silver refrigerator and light brown cabinets .


Predicted Caption: a kitchen with a stove and a stove top oven .
Model: LSTM (baseline)
img_id: 550864
bleu1: 0.646
bleu4: 0.013
```

(e) 4e

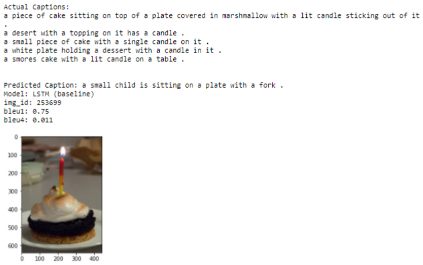
Figure 4: Images and generated captions against actual captions for best model where predictions seemed bad



```
Actual Captions:
a trey with some bread and juice on it
red wine , coffee and pastries are served . there is a statue of a frog riding a snail on the table .
a table has a plate adorned with desert and coffee .
a plate of food with cups on a table
wine and desserts are served on a table .


Predicted Caption: a plate of food with a plate of food
Model: LSTM (baseline)
img_id: 343254
bleu1: 0.667
bleu4: 0.333
```

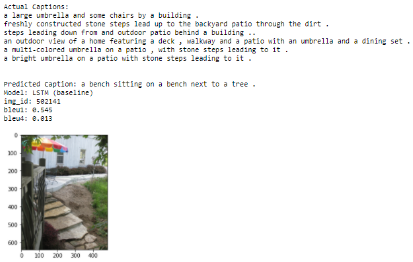
(a) 5a



```
Actual Captions:
a piece of cake sitting on top of a plate covered in marshmallow with a lit candle sticking out of it
.
a desert with a topping on it has a candle .
a small piece of cake with a single candle on it .
a white plate holding a dessert with a candle in it .
a smores cake with a lit candle on a table .


Predicted Caption: a small child is sitting on a plate with a fork .
Model: LSTM (baseline)
img_id: 253699
bleu1: 0.75
bleu4: 0.011
```

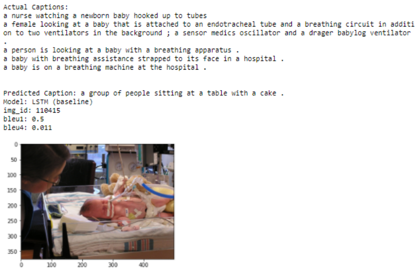
(b) 5b



```
Actual Captions:
a large umbrella and some chairs by a building .
freshly constructed stone steps lead up to the backyard patio through the dirt .
steps leading down from and outdoor patio behind a building ..
an outdoor view of a home featuring a deck , walkway and a patio with an umbrella and a dining set .
a multi-colored umbrella on a patio , with stone steps leading to it .
a bright umbrella on a patio with stone steps leading to it .


Predicted Caption: a bench sitting on a bench next to a tree .
Model: LSTM (baseline)
img_id: 502141
bleu1: 0.545
bleu4: 0.013
```

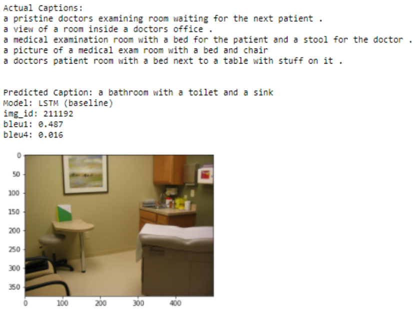
(c) 5c



```
Actual Captions:
a nurse watching a newborn baby hooked up to tubes
a female looking at a baby that is attached to an endotracheal tube and a breathing circuit in additi
on to two ventilators in the background ; a sensor medics oscillator and a drager babylog ventilator
.
a person is looking at a baby with a breathing apparatus .
a baby with breathing assistance strapped to its face in a hospital .
a baby is on a breathing machine at the hospital .


Predicted Caption: a group of people sitting at a table with a cake .
Model: LSTM (baseline)
img_id: 110415
bleu1: 0.5
bleu4: 0.011
```

(d) 5d

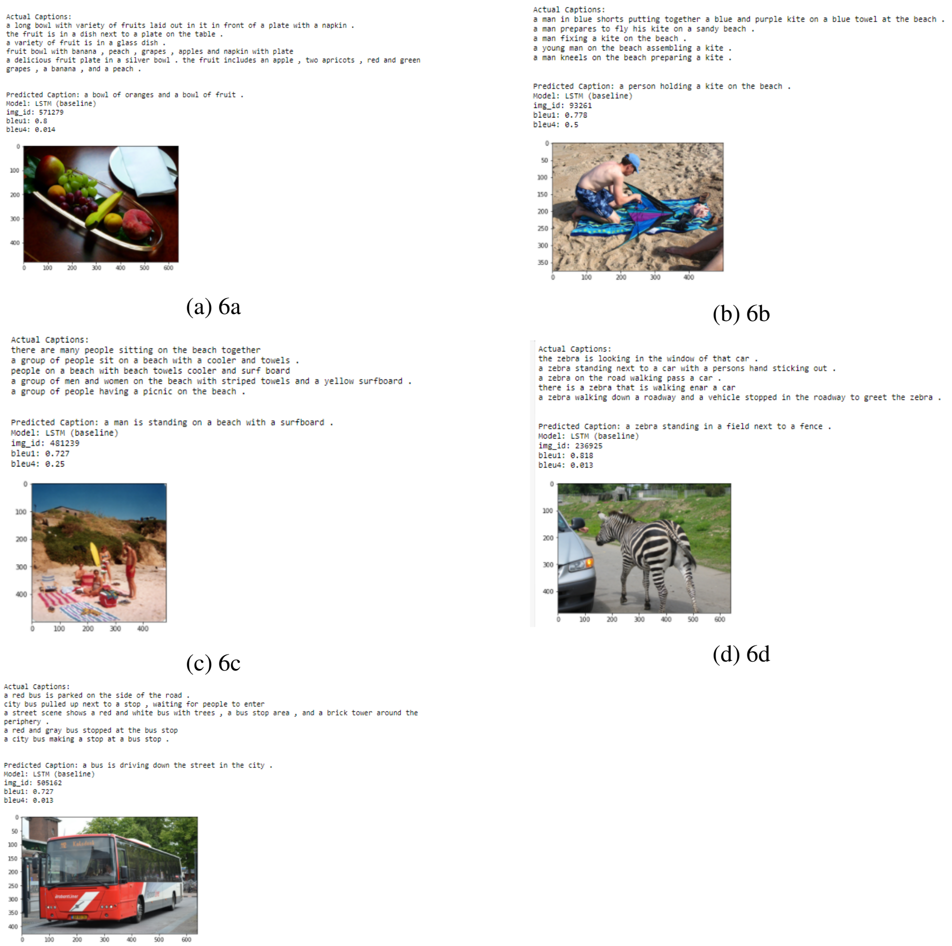


```
Actual Captions:
a pristine doctors examining room waiting for the next patient .
a view of a room inside a doctors office .
a medical examination room with a bed for the patient and a stool for the doctor .
a picture of a medical exam room with a bed and chair
a doctors patient room with a bed next to a table with stuff on it .


Predicted Caption: a bathroom with a toilet and a sink
Model: LSTM (baseline)
img_id: 211192
bleu1: 0.487
bleu4: 0.016
```

(e) 5e

Figure 5: More images and generated captions against actual captions for best model where predictions seemed bad



```
Actual Captions:
a long bowl with variety of fruits laid out in it in front of a plate with a napkin .
the fruit is in a dish next to a plate on the table .
a variety of fruit is in a glass dish .
fruit bowl with banana , peach , grapes , apples and napkin with plate
a delicious fruit plate in a silver bowl . the fruit includes an apple , two apricots , red and green
grapes , a banana , and a peach .


Predicted Caption: a bowl of oranges and a bowl of fruit .
Model: LSTM (baseline)
img_id: 571279
bleu1: 0.8
bleu4: 0.014
```

(a) 6a

```
Actual Captions:
a man in blue shorts putting together a blue and purple kite on a blue towel at the beach .
a man prepares to fly his kite on a sandy beach .
a man fixing a kite on the beach .
a young man on the beach assembling a kite .
a man kneels on the beach preparing a kite .


Predicted Caption: a person holding a kite on the beach .
Model: LSTM (baseline)
img_id: 93261
bleu1: 0.778
bleu4: 0.5
```

(b) 6b

```
Actual Captions:
there are many people sitting on the beach together
a group of people sit on a beach with a cooler and towels .
people on a beach with beach towels cooler and surf board
a group of men and women on the beach with striped towels and a yellow surfboard .
a group of people having a picnic on the beach .


Predicted Caption: a man is standing on a beach with a surfboard .
Model: LSTM (baseline)
img_id: 481239
bleu1: 0.727
bleu4: 0.25
```

(c) 6c

```
Actual Captions:
the zebra is looking in the window of that car .
a zebra standing next to a car with a persons hand sticking out .
a zebra on the road walking pass a car .
there is a zebra that is walking enar a car
a zebra walking down a roadway and a vehicle stopped in the roadway to greet the zebra .


Predicted Caption: a zebra standing in a field next to a fence .
Model: LSTM (baseline)
img_id: 236925
bleu1: 0.818
bleu4: 0.013
```

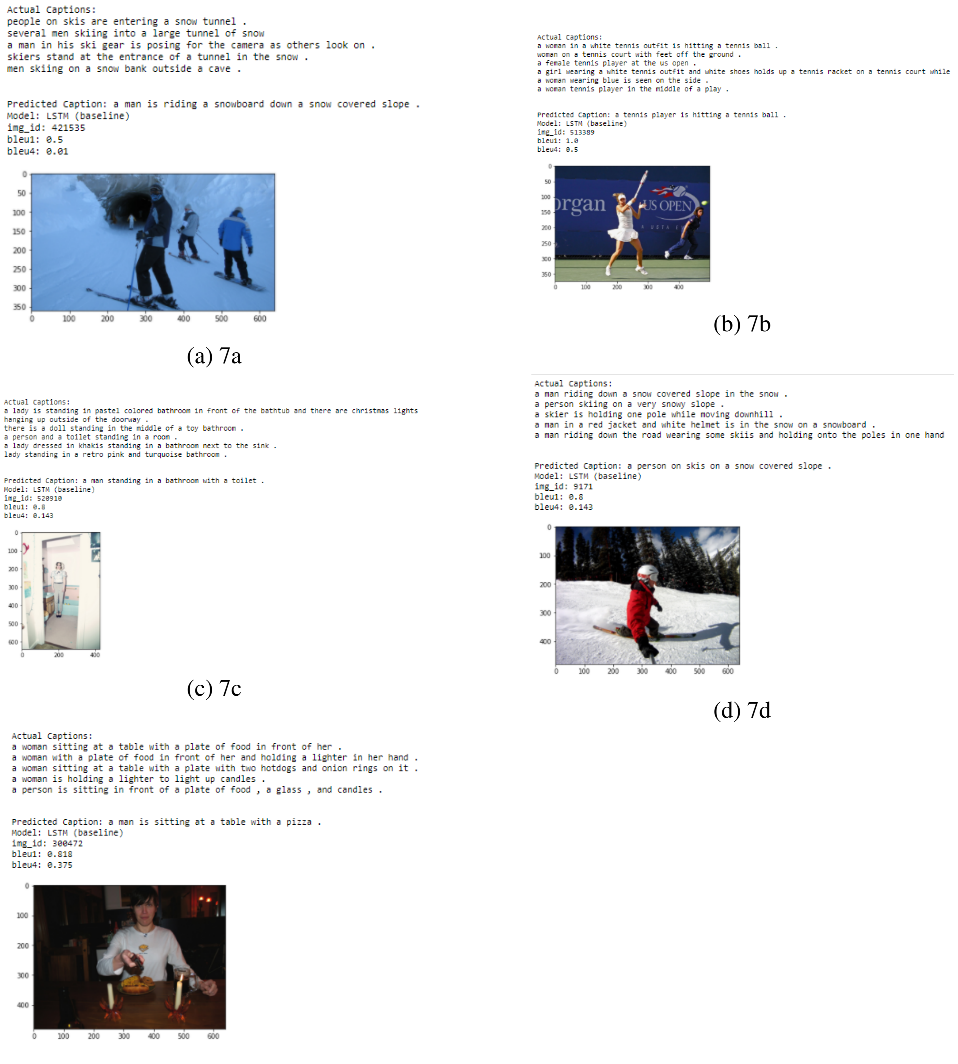
(d) 6d

```
Actual Captions:
a red bus is parked on the side of the road .
city bus pulled up next to a stop , waiting for people to enter
a street scene shows a red and white bus with trees , a bus stop area , and a brick tower around the
periphery .
a red and gray bus stopped at the bus stop
a city bus making a stop at a bus stop .


Predicted Caption: a bus is driving down the street in the city .
Model: LSTM (baseline)
img_id: 505162
bleu1: 0.727
bleu4: 0.013
```

(e) 6e

Figure 6: Images and generated captions against actual captions for best model where predictions seemed good



```
Actual Captions:
people on skis are entering a snow tunnel .
several men skiing into a large tunnel of snow
a man in his ski gear is posing for the camera as others look on .
skiers stand at the entrance of a tunnel in the snow .
men skiing on a snow bank outside a cave .


Predicted Caption: a man is riding a snowboard down a snow covered slope .
Model: LSTM (baseline)
img_id: 421535
bleu1: 0.5
bleu4: 0.01
```

(a) 7a

```
Actual Captions:
a woman in a white tennis outfit is hitting a tennis ball .
woman on a tennis court with feet off the ground .
a female tennis player at the us open .
a girl wearing a white tennis outfit and white shoes holds up a tennis racket on a tennis court while
a woman wearing blue is seen on the side .
a woman tennis player in the middle of a play .


Predicted Caption: a tennis player is hitting a tennis ball .
Model: LSTM (baseline)
img_id: 513389
bleu1: 1.0
bleu4: 0.5
```

(b) 7b

```
Actual Captions:
a lady is standing in pastel colored bathroom in front of the bathtub and there are christmas lights
hanging up outside of the doorway .
there is a doll standing in the middle of a toy bathroom .
a person and a toilet standing in a room .
a lady dressed in khakis standing in a bathroom next to the sink .
lady standing in a retro pink and turquoise bathroom .


Predicted Caption: a man standing in a bathroom with a toilet .
Model: LSTM (baseline)
img_id: 520910
bleu1: 0.8
bleu4: 0.143
```

(c) 7c

```
Actual Captions:
a man riding down a snow covered slope in the snow .
a person skiing on a very snowy slope .
a skier is holding one pole while moving downhill .
a man in a red jacket and white helmet is in the snow on a snowboard .
a man riding down the road wearing some skiis and holding onto the poles in one hand


Predicted Caption: a person on skis on a snow covered slope .
Model: LSTM (baseline)
img_id: 9171
bleu1: 0.8
bleu4: 0.143
```

(d) 7d

```
Actual Captions:
a woman sitting at a table with a plate of food in front of her .
a woman with a plate of food in front of her and holding a lighter in her hand .
a woman sitting at a table with a plate with two hotdogs and onion rings on it .
a woman is holding a lighter to light up candles .
a person is sitting in front of a plate of food , a glass , and candles .


Predicted Caption: a man is sitting at a table with a pizza .
Model: LSTM (baseline)
img_id: 300472
bleu1: 0.818
bleu4: 0.375
```

(e) 7e

Figure 7: More images and generated captions against actual captions for best model where predictions seemed good

10 Team Contributions

10.1 Jacky Dam

In this project, I worked on coding all the different models and figuring out how to calculate the different components: loss, bleu1, bleu4 from the validation/test portion of experiment file. I also gave the team my input on why we got certain results after changing specific parts of the code.

10.2 Dylan Loe

I helped code the stochastic caption generation for our models and helped test out different combinations of hyperparameters for our baseline LSTM model. I helped test our vanilla RNN model and the model with the second architecture. I also wrote parts of sections like the abstract and the introduction. I also helped format all of our images in the paper.

10.3 Ali Zaidi

In this project, I helped in various parts of the coding and debugging, including Architecture 1 and Architecture 2. I also helped with some hyperparameter tuning for the various models.

11 References

[1] Oriol Vinyals and Alexander Toshev and Samy Bengio and Dumitru Erhan (2016). Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning ChallengeCoRR, abs/1609.06647.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, & Piotr Dollár. (2015). Microsoft COCO: Common Objects in Context.

[3] Cottrell, Gary, 2020. Lecture 9, lecture slides, *Recurrent Nets*, UCSD.

[4] Cottrell, Gary, 2020. Lecture 10, lecture slides, *Recurrent Nets Part 2*, UCSD.