Image Captioning Experiments with RNN Models on COCO's dataset

Wenbo Hu, Wenxiao Li, Huaning Liu
University of California, San Diego
w1hu@ucsd.edu, wel032@ucsd.edu, h9liu@ucsd.edu

Abstract

In this project, we are going to design and implement a recurrent neural network that generates captions for images in the provided dataset. This framework contains an encoder (convolutional) that basically takes the images as input, converting them into features; and a decoder (LSTM) that actually predicts and generates the captions. Based on the framework, several related models, including baseline model, Vanilla RNN and Architecture2 are configured, trained, tuned, and tested based on the given data. After hyper-tuning for these three models, related test BLEU-1 and BLEU-4 scores are outputted as the model performance. In this report, related statistics, an in-depth description of methodology and further discussion will be provided. In short, the tuned baseline model with LSTM gives a test BLEU-1 and BLEU-4 score of 70.916 and 6.544; the tuned Vanilla RNN model gives a test BLEU-1 score and BLEU-4 score of 64.5 and 5.5; and the tuned architecture 2 model gives a test BLEU-1 score and BLEU-4 score of 69.474 and 7.270. Further discussion for this result follows up.

1 General Introduction

Generating captions based on images is a general practice for deep learning on computer vision. In the real world, when we are to search text by images or search images by images, texts take less time for the engine to process, and it also better represents what people are trying to express. Therefore, converting images to text that describe and represent it will be meaningful given such potential applications, and it would become the main problem for this project. To build a good network for this, choosing good trainable datasets and proper network architecture are both important. To tackle this, the dataset we are going to utilize this time comes from COCO 2015 Image Captioning Task, where COCO is a "large-scale detection, segmentation, and captioning dataset". In this dataset, for train and test data respectively, about 82k and 3k images with each respect to several captions are provided. Given such a large size of data and the upcoming complex architectures, we only use $\frac{1}{5}$ of it. Also, to better evaluate such image captioning models, different from the previous project, we are going to use BLEU (bilingual evaluation understudy) score as the evaluation metrics. This score actually represents the quality of text which has been machine-translated from the natural language (images here) to another. It actually takes a very high correlation with human judgments of text quality. Accordingly, the BLEU-n score refers to the precision of text prediction under n-gram. In this case, both BLEU-1 and BLEU-4 are presented to better evaluate our models. In this way, with the dataset and metrics getting well-prepared, we get started on the framework and its related models.

2 Related Works

Since this dataset comes from COCO 2015 Image Captioning Task, several reflection articles and papers emerge to address this topic from different perspectives, including hidden image captioning,

better text generations, etc. Based on this project's problems and planning works this time, Oriol Vinyals and his team's work, **Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge** [1], gives us a good reference. In this paper, based on COCO 2015 image captioning dataset, the team built a deep learn CNN as well as language generating RNN together as the overall architecture. Furthermore, other related experiments regarding hyper-tuning and different approaches are provided and analyzed. Besides, extended from our planning, they also discuss the possibilities of using different evaluation metrics, such as METEOR and CIDER. This work gives us a matching initiative, whose ideas and sub-methodologies will definitely be discussed to contribute to our work.

3 Methods

3.1 LSTM Model (from Baseline)

For the baseline LSTM model, it first uses a pretrained convolutional network (resnet50) as the encoder the get the feature vectors, then construct a network with consecutive LSTM cells as the decoder that take the feature vectors and generate a sentence of words signified by starting with the word "⟨ start ⟩" and ending by "⟨ end ⟩". The feature vectors of the image will be passed to the decoder initially, generating the ⟨ start ⟩ symbol, and a hidden state. Then, for the following LSTMs, the hidden state of the last LSTM, together with the embedding of the last word generated, will be passed in. nn.Embeddings is here used to address this problem by configuring a one-hot encoding as well as linear output. This procedure stops when reaching max length, or the ⟨ end ⟩ symbol is generated. The outputs from the decoder are sampled and generated using greedy search, which also depends on the value of deterministic. If it goes true, the word directly takes the one with the largest; while for false, a probabilistic series is generated by softmax to decide each word. For the hyper-tuning process, we focus on the adjustment of the embedding size, and hidden size. The embedding size is adjusted within the range of 250 and 750, and the hidden size is adjusted within the range of 256 and 1024. After tuning, the best parameters we find for the baseline LSTM model is settled as below.

3.2 Vanilla RNN

The overall architecture and logic of the Vanilla RNN model follow the pattern of the baseline model with LSTM. It first uses a pretrained convolutional network (resnet50) as the encoder the get the feature vectors, then construct a network with RNN (trained by BPTT) cells as the decoder that take the feature vectors and generate a sentence of words signified by starting with the word "⟨ start ⟩" and ending by "⟨ end ⟩". nn.Embeddings is here used to address this problem by configuring a one-hot encoding as well as linear output. The outputs from the decoder are sampled and generated using greedy search, which also depends on the value of deterministic. If it goes true, the word directly takes the one with the largest; while for false, a probabilistic series is generated by softmax to decide each word. For the hyper-tuning process, we focus on the adjustment of the optimizer, embedding size, and hidden size. The optimizer has candidates of SGD and Adam, the embedding size is adjusted within the range of 250 and 750, and the hidden size is adjusted within the range of 256 and 1024. As usual, we found that when parameters go extreme, overfitting happens. After tuning, the best parameters we find for the Vanilla RNN model is settled as below.

3.3 Arch 2

For architecture 2, it first uses a pretrained convolutional network (resnet50) as the encoder the get the feature vectors, then construct a network with consecutive LSTM cells as the decoder that take the feature vectors and generate a sentence of words signified by starting with the word "⟨ start ⟩" and ending by "⟨ end ⟩". The feature vectors of the image, concatenated with the word embedding of the ⟨ pad ⟩ symbol, will be passed to the decoder initially, generating the ⟨ start ⟩ symbol, and a hidden state. Then, for the following LSTMs, the hidden state of the last LSTM, and feature vectors of the image, concatenated with the embedding of the last word generated, will be passed in. This procedure stops when reaching max length, or the ⟨ end ⟩ symbol is generated. The outputs from the decoder are sampled and generated using greedy search, which also depends on the value of deterministic. If it goes true, the word directly takes the one with the largest; while

for false, a probabilistic series is generated by softmax to decide each word. The hyperparameters of architecture 2 are set to be the same as the baseline LSTM model, for better observations of the impact caused by difference in architectures.

4 Results: Best Hyperparameters

4.1 LSTM Model (from Baseline)

The best hyperparameter we achieved for the Vanilla RNN model will be: for experiment-wise, the number of epochs is settled to be 10, with a learning rate of 5e-4. For model-wise, the hidden size is adjusted to be 1024, with an embedding size of 500.

4.2 Vanilla RNN

The best hyperparameter we achieved for the Vanilla RNN model will be: for experiment-wise, the number of epochs is settled to be 10, with a learning rate of 5e-4. For model-wise, the hidden size is adjusted to be 512, with an embedding size of 550.

4.3 Arch 2

To better compare Architecture 2 to the baseline LSTM model, we stick all the hyperparameters to the best hyperparameters of the baseline LSTM model: the number of epochs is settled to be 10, with a learning rate of 5e-4. For model-wise, the hidden size is 1024, with an embedding size of 1000(500 for image features, 500 for word embeddings).

5 Results: Visualization and Output Report

5.1 LSTM Model (from Baseline)

The plot 1 shows the training loss and validation loss vs the number of epochs for the baseline LSTM model. Accordingly, the cross-entropy loss on the test set is reported to be about 2.5210.

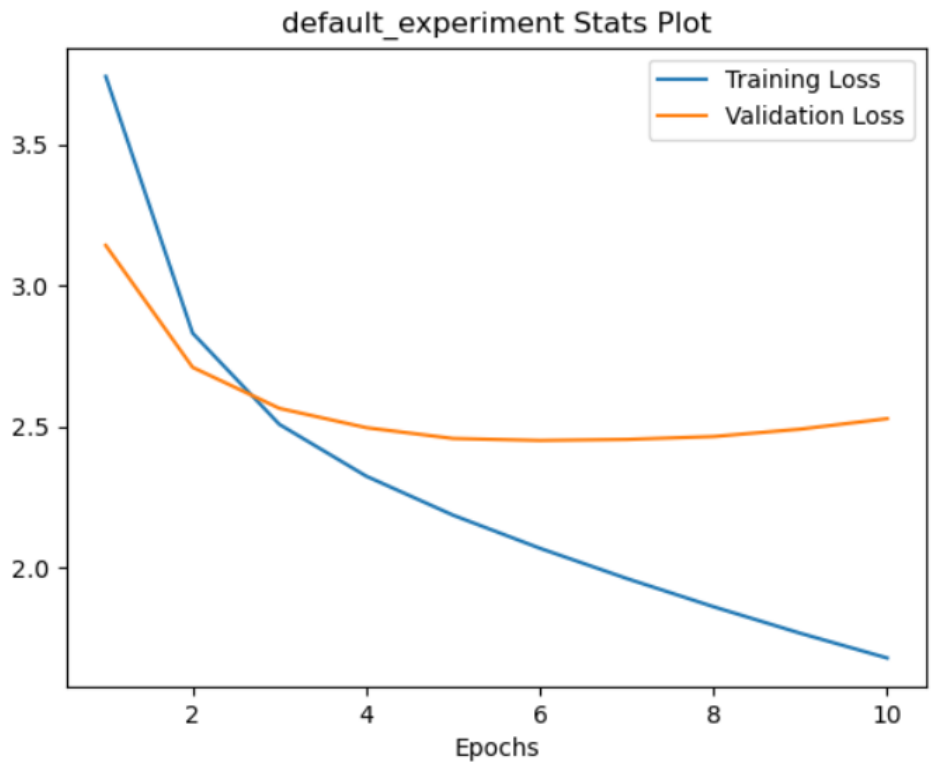


Figure 1: Baseline(LSTM) Model Loss Plot

5.2 Vanilla RNN

The plot 2 shows the training loss and validation loss vs the number of epochs for the Vanilla RNN model. Accordingly, the cross-entropy loss on the test set is reported to be about 2.49.

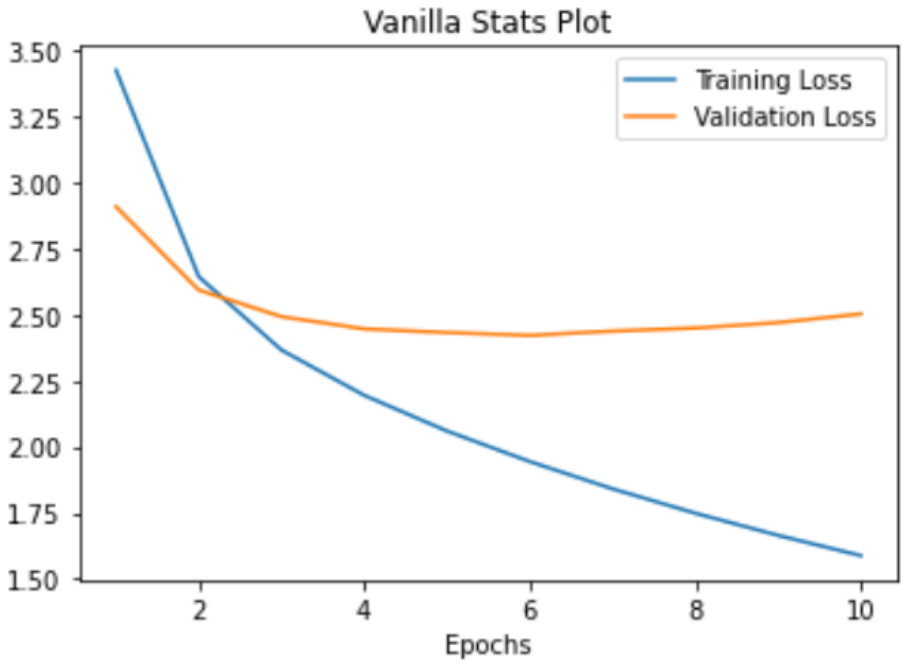


Figure 2: Vanilla RNN Model Loss Plot

5.3 Arch 2

The plot 3 shows the training loss and validation loss vs the number of epochs for the architecture 2. Accordingly, the cross-entropy loss on the test set is reported to be about 2.4545.

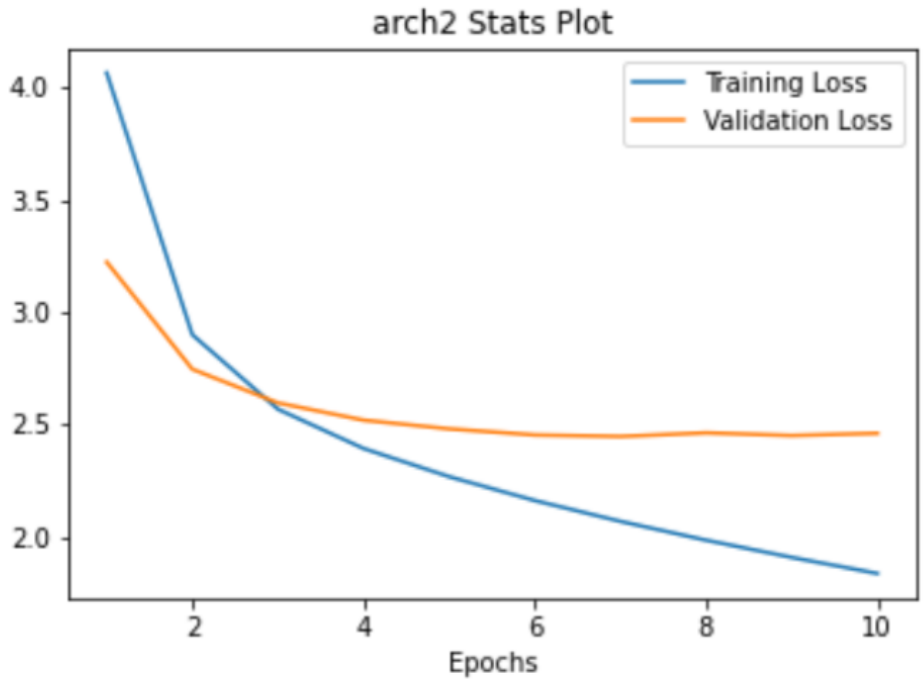


Figure 3: Arch 2 Model Loss Plot

6 Results: BLEU Scores Report

6.1 LSTM Model (from Baseline)

For the LSTM model with default parameters (hidden size = 300, embedding size = 512), the BLEU-1 score on the test set is reported to be about 66.577, and the BLEU-4 score on the test set is reported to be about 5.636. For the LSTM model with tuned parameters (hidden size = 500, embedding size = 1024), the BLEU-1 score on the test set is reported to be about 70.916, and the BLEU-4 score on the test set is reported to be about 6.544.

6.2 Vanilla RNN

For the tuned Vanilla RNN model, the BLEU-1 score on the test set is reported to be about 64.5, and the BLEU-4 score on the test set is reported to be about 5.5.

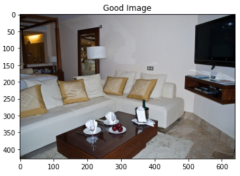
6.3 Arch 2

For the LSTM model with tuned parameters (hidden size = 500, embedding size = 1024), the BLEU-1 score on the test set is reported to be about 69.474, and the BLEU-4 score on the test set is reported to be about 7.270.

7 Image Caption Examples

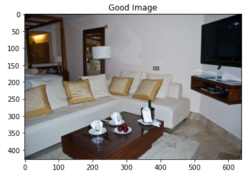
7.1 LSTM Model (from Baseline)

With temperature default = 0.4:



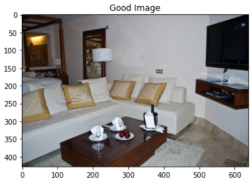
Actual captions: ['A L shaped couch with a variety of pillows on it, in front of a television. ', 'A bottle of wine, place settings, and some apples on a coffee table in front of a couch.', 'THERE IS A PHOTO OF A LIVING ROOM WITH COUCH AND TV', 'a living room filled with couches, tv, coffee table, and a door way', 'a white couch tan pillows coffee table and a television set']
Predicted sentence: a room with a couch , coffee table , and a television .
bleu1 score: 92.3076923076923

With temperature = 5



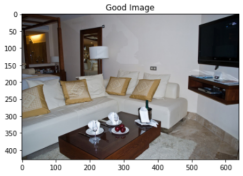
Actual captions: ['A L shaped couch with a variety of pillows on it, in front of a television. ', 'A bottle of wine, place settings, and some apples on a coffee table in front of a couch.', 'THERE IS A PHOTO OF A LIVING ROOM WITH COUCH AND TV', 'a living room filled with couches, tv, coffee table, and a door way', 'a white couch tan pillows coffee table and a television set']
Predicted sentence: royal makeshift durham couple aims newlywed quilt also coronation flattened shapes nicely fish fancy square elegant colored patriotic
bleu1 score: 0

With temperature = 0.001



Actual captions: ['A L shaped couch with a variety of pillows on it, in front of a television. ', 'A bottle of wine, place settings, and some apples on a coffee table in front of a couch.', 'THERE IS A PHOTO OF A LIVING ROOM WITH COUCH AND TV', 'a living room filled with couches, tv, coffee table, and a door way', 'a white couch tan pillows coffee table and a television set']
Predicted sentence: a living room with a couch and a television .
bleu1 score: 81.43536762323636

With temperature deterministic



Actual captions: ['A L shaped couch with a variety of pillows on it, in front of a television. ', 'A bottle of wine, place settings, and some apples on a coffee table in front of a couch.', 'THERE IS A PHOTO OF A LIVING ROOM WITH COUCH AND TV', 'a living room filled with couches, tv, coffee table, and a door way', 'a white couch tan pillows coffee table and a television set']
Predicted sentence: a living room with a couch and a television .
bleu1 score: 81.43536762323636

Figure 4: Good Caption 1

With temperature default = 0.4:

Actual captions: ['a bathroom that has a sink and a toilet in it', 'A bathroom with marble counter top and red wall treatment.', 'A bathroom with a red color scheme on the walls.', 'A bathroom that has a tub, a toilet, a sink, and a mirror. ', 'A bathroom with a tub, toilet, sink, a counter and a cabinet in it.']
Predicted sentence: a bathroom with a toilet , sink and mirror in it .
bleu1 score: 100.0

With temperature = 5

Actual captions: ['a bathroom that has a sink and a toilet in it', 'A bathroom with marble counter top and red wall treatment.', 'A bathroom with a red color scheme on the walls.', 'A bathroom that has a tub, a toilet, a sink, and a mirror. ', 'A bathroom with a tub, toilet, sink, a counter and a cabinet in it.']
Predicted sentence: accessories curtains man given headset customers instructors about ballplayer snapshots plain walking clothing lloyd ramp officials and stove
bleu1 score: 5.555555555555554

With temperature = 0.001

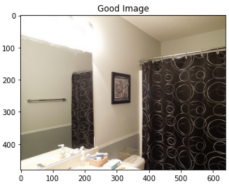
Actual captions: ['a bathroom that has a sink and a toilet in it', 'A bathroom with marble counter top and red wall treatment.', 'A bathroom with a red color scheme on the walls.', 'A bathroom that has a tub, a toilet, a sink, and a mirror. ', 'A bathroom with a tub, toilet, sink, a counter and a cabinet in it.']
Predicted sentence: a bathroom with a toilet , sink and shower .
bleu1 score: 81.43536762323636

With temperature deterministic

Actual captions: ['a bathroom that has a sink and a toilet in it', 'A bathroom with marble counter top and red wall treatment.', 'A bathroom with a red color scheme on the walls.', 'A bathroom that has a tub, a toilet, a sink, and a mirror. ', 'A bathroom with a tub, toilet, sink, a counter and a cabinet in it.']
Predicted sentence: a bathroom with a toilet , sink and shower .
bleu1 score: 81.43536762323636

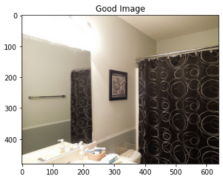
Figure 5: Good Caption 2

With temperature default = 0.4:



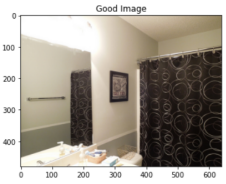
Actual captions: ['A bathroom with a shower covered in a curtain next to a toilet.', 'A country white bathroom takes on an entirely new fashion.', 'A bathroom with a large square mirror over the sink and a brown shower curtain with circle designs.', 'A shower curtain is covering the tub in front of the sink. ', 'Bathroom with a black and white shower curtain']
Predicted sentence: a bathroom with a sink and a toilet
bleu1 score: 100.0

With temperature = 0.001



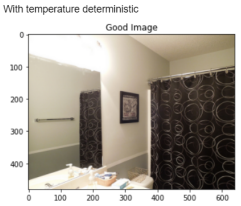
Actual captions: ['A bathroom with a shower covered in a curtain next to a toilet.', 'A country white bathroom takes on an entirely new fashion.', 'A bathroom with a large square mirror over the sink and a brown shower curtain with circle designs.', 'A shower curtain is covering the tub in front of the sink. ', 'Bathroom with a black and white shower curtain']
Predicted sentence: a bathroom with a toilet , sink and shower .
bleu1 score: 81.43536762323636

With temperature = 5



Actual captions: ['A bathroom with a shower covered in a curtain next to a toilet.', 'A country white bathroom takes on an entirely new fashion.', 'A bathroom with a large square mirror over the sink and a brown shower curtain with circle designs.', 'A shower curtain is covering the tub in front of the sink. ', 'Bathroom with a black and white shower curtain']
Predicted sentence: neon up-close photo electricity environment seat wish government debris road shire czech overpass fencing bushes framed prepare grasps
bleu1 score: 0

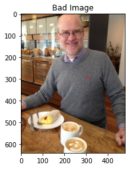
With temperature deterministic



Actual captions: ['A bathroom with a shower covered in a curtain next to a toilet.', 'A country white bathroom takes on an entirely new fashion.', 'A bathroom with a large square mirror over the sink and a brown shower curtain with circle designs.', 'A shower curtain is covering the tub in front of the sink. ', 'Bathroom with a black and white shower curtain']
Predicted sentence: a bathroom with a toilet , sink and shower .
bleu1 score: 81.43536762323636

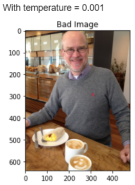
Figure 6: Good Caption 3

With temperature default = 0.4:



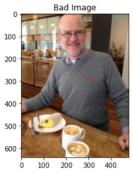
Actual captions: ['A man smiling with a breakfast burrito and a coffee cup with a smiley face.', 'A smiling man stands at a table with an egg burrito and two mugs of coffee.', 'a person at a table with a plate of food and a cup', 'A man sitting at a table with a burrito and a cup of coffee in front of him.', 'A man in grey sweater standing at a wooden table with food and coffee.']
Predicted sentence: a man is eating a piece of chocolate cake .
bleu1 score: 37.040911034085894

With temperature = 0.001



Actual captions: ['A man smiling with a breakfast burrito and a coffee cup with a smiley face.', 'A smiling man stands at a table with an egg burrito and two mugs of coffee.', 'a person at a table with a plate of food and a cup', 'A man sitting at a table with a burrito and a cup of coffee in front of him.', 'A man in grey sweater standing at a wooden table with food and coffee.']
Predicted sentence: a man is sitting at a table with a plate of food .
bleu1 score: 92.3076923076923

With temperature = 5



Actual captions: ['A man smiling with a breakfast burrito and a coffee cup with a smiley face.', 'A smiling man stands at a table with an egg burrito and two mugs of coffee.', 'a person at a table with a plate of food and a cup', 'A man sitting at a table with a burrito and a cup of coffee in front of him.', 'A man in grey sweater standing at a wooden table with food and coffee.']
Predicted sentence: aeroplanes woman total poster corner suggesting sold hog rod . coned roofed woman like newspaper tractor 16 fields
bleu1 score: 5.555555555555554

With temperature deterministic

Actual captions: ['A man smiling with a breakfast burrito and a coffee cup with a smiley face.', 'A smiling man stands at a table with an egg burrito and two mugs of coffee.', 'a person at a table with a plate of food and a cup', 'A man sitting at a table with a burrito and a cup of coffee in front of him.', 'A man in grey sweater standing at a wooden table with food and coffee.']
Predicted sentence: a man is sitting at a table with a plate of food .
bleu1 score: 92.3076923076923

Figure 7: Bad Caption 1

With temperature default = 0.4:

Actual captions: ['An old tour bus for Buddy Holly and the Crickets', 'A large red bus parked in a stationary position.', 'An old fashioned red and white tour bus with the enscription "Buddy Holly and the Crickets."', 'A tour bus advertises Buddy Holly and Ritchie Valens. ', 'A red and white bus on street with trees in the background.']
Predicted sentence: a fire hydrant with a yellow top and a red tail on the side of a building .
bleu1 score: 36.78731267970755

With temperature = 5

Actual captions: ['An old tour bus for Buddy Holly and the Crickets', 'A large red bus parked in a stationary position.', 'An old fashioned red and white tour bus with the enscription "Buddy Holly and the Crickets."', 'A tour bus advertises Buddy Holly and Ritchie Valens. ', 'A red and white bus on street with trees in the background.']
Predicted sentence: pants cells echo parasail interact outdoor rags hiding appears that stands before virgin explains search hovers boot smoke
bleu1 score: 0

With temperature = 0.001

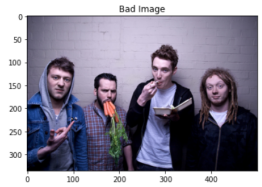
Actual captions: ['An old tour bus for Buddy Holly and the Crickets', 'A large red bus parked in a stationary position.', 'An old fashioned red and white tour bus with the enscription "Buddy Holly and the Crickets."', 'A tour bus advertises Buddy Holly and Ritchie Valens. ', 'A red and white bus on street with trees in the background.']
Predicted sentence: a red double decker bus driving down the street .
bleu1 score: 60.0

With temperature deterministic

Actual captions: ['An old tour bus for Buddy Holly and the Crickets', 'A large red bus parked in a stationary position.', 'An old fashioned red and white tour bus with the enscription "Buddy Holly and the Crickets."', 'A tour bus advertises Buddy Holly and Ritchie Valens. ', 'A red and white bus on street with trees in the background.']
Predicted sentence: a red double decker bus driving down the street .
bleu1 score: 60.0

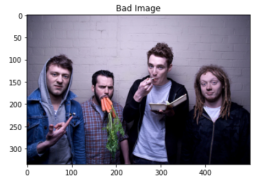
Figure 8: Bad Caption 2

With temperature default = 0.4:



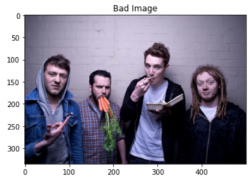
Actual captions: ['A couple of people that are posing for a picture.', 'Four men standing against a brick wall with one of them eating from a food container, one with several carrots stuffed into him mouth and the other two standing next to them.', 'Four men standing, one with an entire bunch of carrots in his mouth.', 'A group of men posing together in various ways.', 'A guy points at a guy with carrots in his mouth. ']
Predicted sentence: a man is holding up a large sandwich
bleu1 score: 19.470019576785123

With temperature = 5



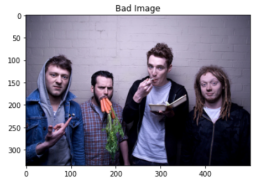
Actual captions: ['A couple of people that are posing for a picture.', 'Four men standing against a brick wall with one of them eating from a food container, one with several carrots stuffed into him mouth and the other two standing next to them.', 'Four men standing, one with an entire bunch of carrots in his mouth.', 'A group of men posing together in various ways.', 'A guy points at a guy with carrots in his mouth. ']
Predicted sentence: join changing safari driven named reception ski village provided kearny clocks ears railroad shes jal grilled checks preserved
bleu1 score: 0

With temperature = 0.001



Actual captions: ['A couple of people that are posing for a picture.', 'Four men standing against a brick wall with one of them eating from a food container, one with several carrots stuffed into him mouth and the other two standing next to them.', 'Four men standing, one with an entire bunch of carrots in his mouth.', 'A group of men posing together in various ways.', 'A guy points at a guy with carrots in his mouth. ']
Predicted sentence: a man and a woman are sitting together
bleu1 score: 48.67504894196281

With temperature deterministic



Actual captions: ['A couple of people that are posing for a picture.', 'Four men standing against a brick wall with one of them eating from a food container, one with several carrots stuffed into him mouth and the other two standing next to them.', 'Four men standing, one with an entire bunch of carrots in his mouth.', 'A group of men posing together in various ways.', 'A guy points at a guy with carrots in his mouth. ']
Predicted sentence: a man and a woman are sitting together
bleu1 score: 48.67504894196281

Figure 9: Bad Caption 3

7.2 Vanilla RNN

Figure 10: Good Caption 1

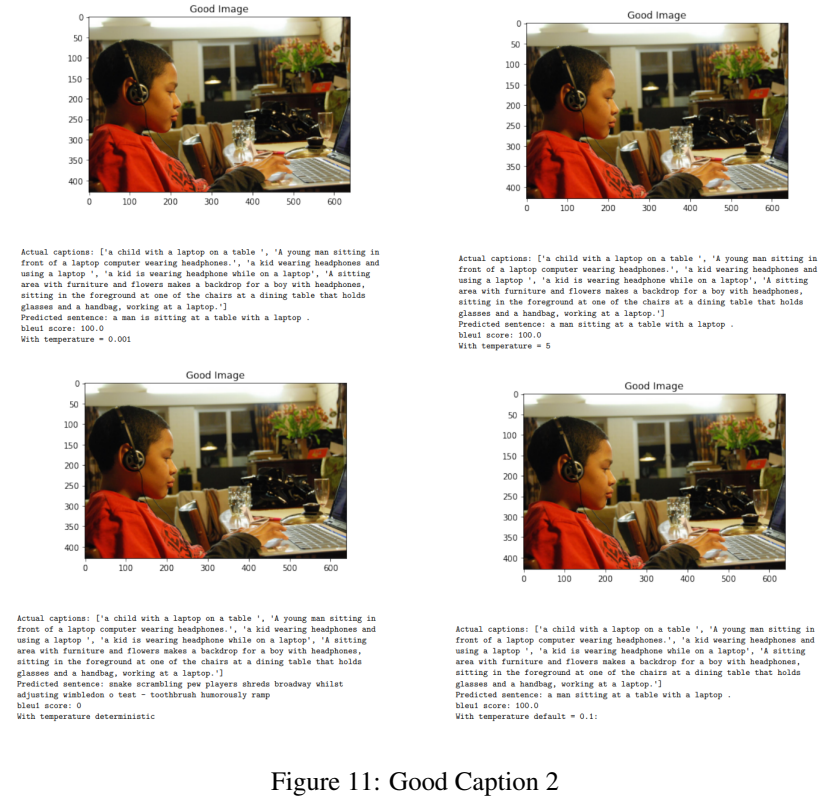


Figure 11: Good Caption 2

Figure 12: Good Caption 3

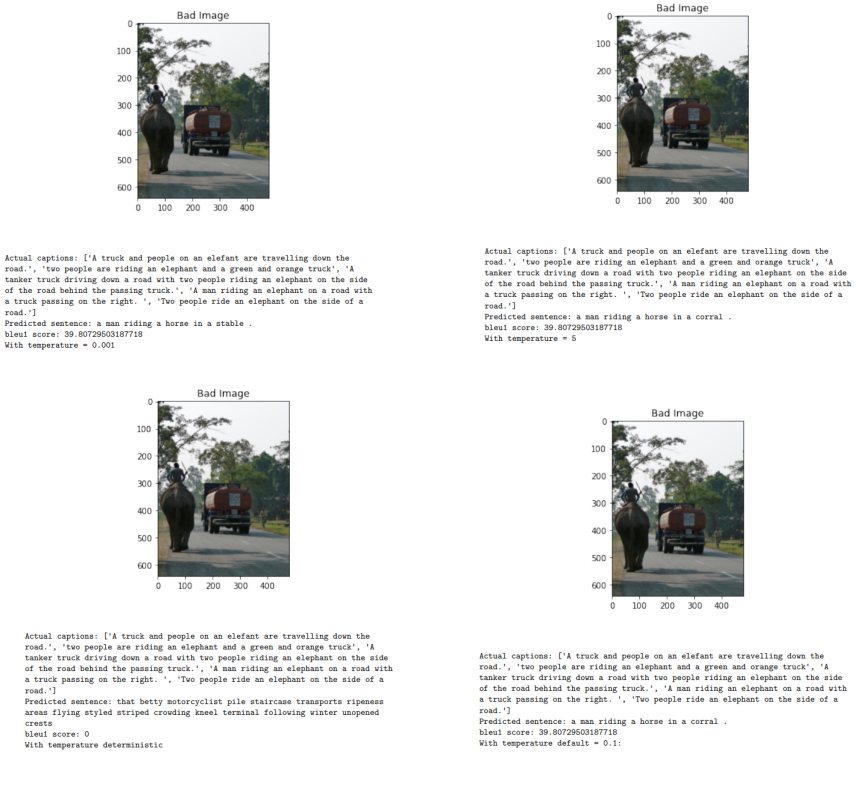


Figure 13: Bad Caption 1

Figure 14: Bad Caption 2

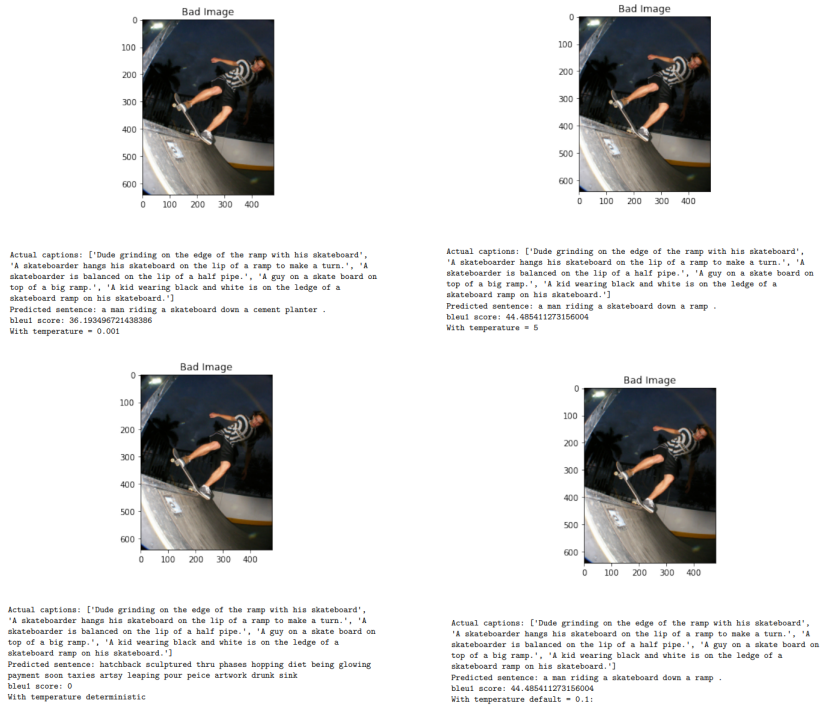


Figure 15: Bad Caption 3

7.3 Arch 2

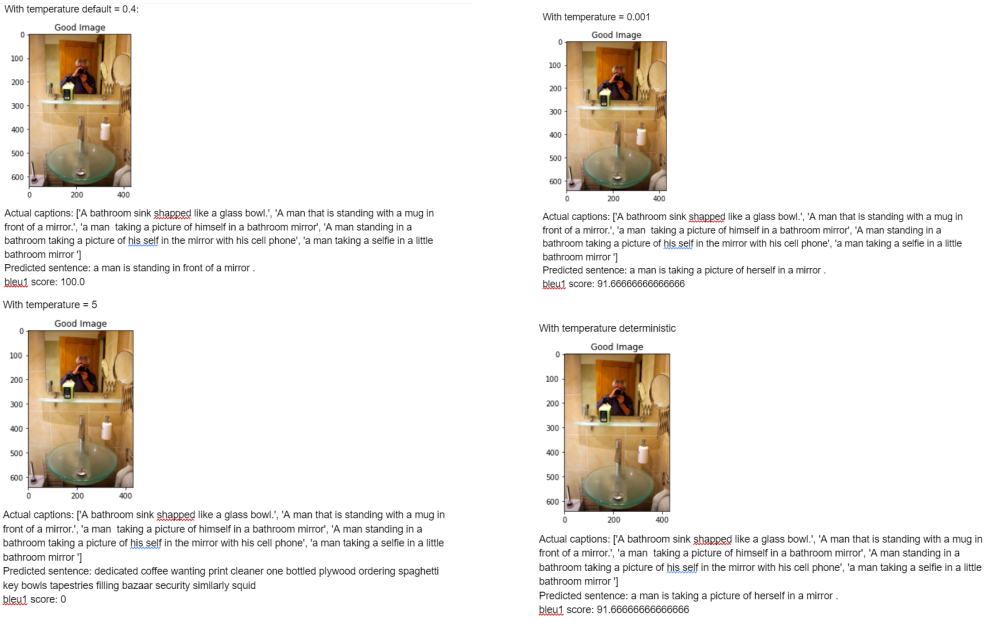
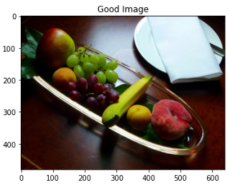


Figure 16: Good Caption 1

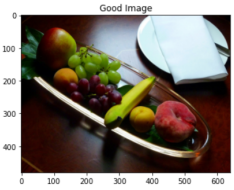
Figure 17: Good Caption 2

With temperature default = 0.4:



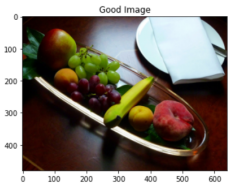
Actual captions: ['A long bowl with variety of fruits laid out in it in front of a plate with a napkin.', 'The fruit is in a dish next to a plate on the table.', 'A variety of fruit is in a glass dish.', 'Fruit bowl with banana, peach, grapes, apples and napkin with plate', 'A delicious fruit plate in a silver bowl. The fruit includes an apple, two apricots, red and green grapes, a banana, and a peach.']
Predicted sentence: a bowl of fruit and a banana sitting on a table .
bleu1 score: 91.66666666666666

With temperature = 5



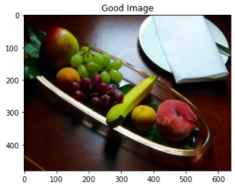
Actual captions: ['A long bowl with variety of fruits laid out in it in front of a plate with a napkin.', 'The fruit is in a dish next to a plate on the table.', 'A variety of fruit is in a glass dish.', 'Fruit bowl with banana, peach, grapes, apples and napkin with plate', 'A delicious fruit plate in a silver bowl. The fruit includes an apple, two apricots, red and green grapes, a banana, and a peach.']
Predicted sentence: boa spear kitty snoozes cherries snacking getting gauze multicolored skin spiral television toyota soft marked tails tufts "
bleu1 score: 0

With temperature = 0.001



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 sentence: a bowl of fruit and a banana sitting on a table .
bleu1 score: 91.66666666666666

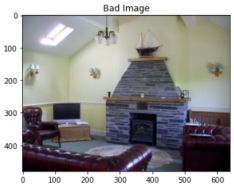
With temperature deterministic



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 sentence: a bowl of fruit and a banana sitting on a table .
bleu1 score: 91.66666666666666

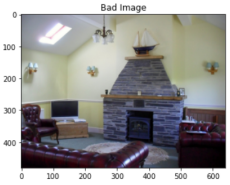
Figure 18: Good Caption 3

With temperature default = 0.4:



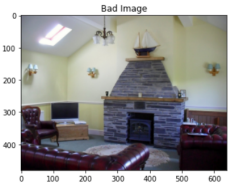
Actual captions: ['A couch and a chair in a small room.', 'a cozy room with leather furniture and a fireplace', 'a fireplace a red couch two chairs and some lights', 'A living room has plush furniture and a fireplace.', 'This living room has leather furniture, a television, and a model ship on top of a fireplace. ']
Predicted sentence: a room with a couch , a couch , a couch , a couch , a couch ,
bleu1 score: 39.77063630286087

With temperature = 5



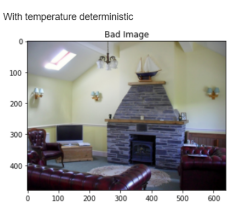
Actual captions: ['A couch and a chair in a small room.', 'a cozy room with leather furniture and a fireplace', 'a fireplace a red couch two chairs and some lights', 'A living room has plush furniture and a fireplace.', 'This living room has leather furniture, a television, and a model ship on top of a fireplace. ']
Predicted sentence: books swift fillings tomato icicles blowing taller frowns curio booth featuring heavily prowl frumpled decorations stop windows times
bleu1 score: 0

With temperature = 0.001



Actual captions: ['A couch and a chair in a small room.', 'a cozy room with leather furniture and a fireplace', 'a fireplace a red couch two chairs and some lights', 'A living room has plush furniture and a fireplace.', 'This living room has leather furniture, a television, and a model ship on top of a fireplace. ']
Predicted sentence: a room with a couch , a couch , a couch , a couch , a couch ,
bleu1 score: 39.77063630286087

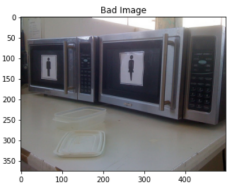
With temperature deterministic



Actual captions: ['A couch and a chair in a small room.', 'a cozy room with leather furniture and a fireplace', 'a fireplace a red couch two chairs and some lights', 'A living room has plush furniture and a fireplace.', 'This living room has leather furniture, a television, and a model ship on top of a fireplace. ']
Predicted sentence: a room with a couch , a couch , a couch , a couch , a couch ,
bleu1 score: 39.77063630286087

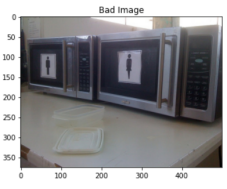
Figure 19: Bad Caption 1

With temperature default = 0.4:



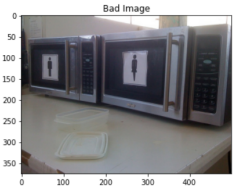
Actual captions: ['Two microwave ovens one marked with man symbol and the other for woman. ', 'Two microwaves sitting side by side on a countertop are marked with signs printed with the symbols for man and woman.', 'a couple of microwaves that are next to each other', 'Two microwaves side by side with pictures of a man and woman silhouette', 'The microwaves are marked with female and male signs.']
Predicted sentence: a kitchen with a stove and a stove
bleu1 score: 29.20502936517768

With temperature = 0.001



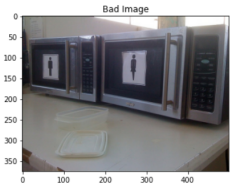
Actual captions: ['Two microwave ovens one marked with man symbol and the other for woman. ', 'Two microwaves sitting side by side on a countertop are marked with signs printed with the symbols for man and woman.', 'a couple of microwaves that are next to each other', 'Two microwaves side by side with pictures of a man and woman silhouette', 'The microwaves are marked with female and male signs.']
Predicted sentence: a kitchen with a stove and a stove
bleu1 score: 29.20502936517768

With temperature = 5



Actual captions: ['Two microwave ovens one marked with man symbol and the other for woman. ', 'Two microwaves sitting side by side on a countertop are marked with signs printed with the symbols for man and woman.', 'a couple of microwaves that are next to each other', 'Two microwaves side by side with pictures of a man and woman silhouette', 'The microwaves are marked with female and male signs.']
Predicted sentence: jeep nicely coverlet multiple car guards takeoff sign shelf specially managers cleaning fives suited suit noah gone low
bleu1 score: 0

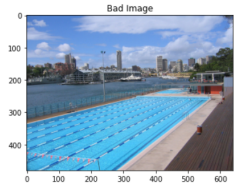
With temperature deterministic



Actual captions: ['Two microwave ovens one marked with man symbol and the other for woman. ', 'Two microwaves sitting side by side on a countertop are marked with signs printed with the symbols for man and woman.', 'a couple of microwaves that are next to each other', 'Two microwaves side by side with pictures of a man and woman silhouette', 'The microwaves are marked with female and male signs.']
Predicted sentence: a kitchen with a stove and a stove
bleu1 score: 29.20502936517768

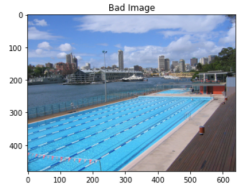
Figure 20: Bad Caption 2

With temperature default = 0.4:



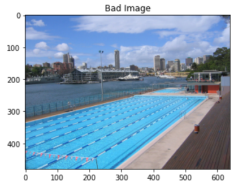
Actual captions: ["A swimming pool with lanes marked for racing on the water's edge across from a large city.", 'An Olympic style swimming pool, with many dividers', 'an Olympic pool sitting next to a river ', 'A large swimming pool is next to a large harbor.', 'A swimming pool that is next to a ocean.']
Predicted sentence: a bench and a bench on a sunny day
bleu1 score: 33.333333333333333

With temperature = 0.001



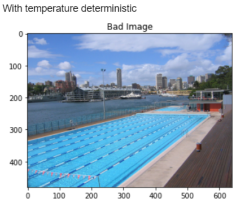
Actual captions: ["A swimming pool with lanes marked for racing on the water's edge across from a large city.", 'An Olympic style swimming pool, with many dividers', 'an Olympic pool sitting next to a river ', 'A large swimming pool is next to a large harbor.', 'A swimming pool that is next to a ocean.']
Predicted sentence: a bench and a bench on a beach
bleu1 score: 37.5

With temperature = 5



Actual captions: ["A swimming pool with lanes marked for racing on the water's edge across from a large city.", 'An Olympic style swimming pool, with many dividers', 'an Olympic pool sitting next to a river ', 'A large swimming pool is next to a large harbor.', 'A swimming pool that is next to a ocean.']
Predicted sentence: sundown occupant insignia smaller rodeo w walkway showing quarters scape chateau sails accord upon banner message stand moored
bleu1 score: 0

With temperature deterministic



Actual captions: ["A swimming pool with lanes marked for racing on the water's edge across from a large city.", 'An Olympic style swimming pool, with many dividers', 'an Olympic pool sitting next to a river ', 'A large swimming pool is next to a large harbor.', 'A swimming pool that is next to a ocean.']
Predicted sentence: a bench and a bench on a beach
bleu1 score: 37.5

Figure 21: Bad Caption 3

8 Discussion and Findings

8.1 Why does the deterministic approach not work well?

Deterministic approach only takes the maximum probability of a word at each step. While for stochastic processes, setting a temperature parameter makes the modeler control the stochastic attribute of the distribution. If temperature is close to 0, then the distribution is nearly the same as the deterministic approach, if the temperature is approaching infinity, then the distribution is almost

a uniform distribution. So it gives a chance of all the output words according to their probability distribution. The deterministic approach selects the most confident word. However, sometimes we don't want the next word only to be very focused on the previous word, we want the next word to also consider image or the whole correlation in the sentence which could be the second or the third most confident word. Thus, the stochastic approach is better and more reasonable.

8.2 LSTM Model (from Baseline)

From the results of the baseline models, we are able to observe that choosing a proper temperature to generate captions is crucial. When the temperature is too big, such as 5, words will be selected in a roughly uniform distribution, and the predictions do not make any sense. When temperature is very small, such as 0.001, it is almost certain that the words with the largest output value will be selected. From the outputs shown, we can see that the output captions for deterministic settings are the same as the output captions when temperature equals 0.001. There are advantages for adding a little bit of uncertainty to generate captions by introducing a temperature such as 0.4. For instance, in the first good picture of the bathroom, it is able to say there is a mirror. However, when it is set to deterministic, both scenarios will have output "a bathroom with a toilet, sink and shower." This is probably due to the inherent connection between bathroom, toilet, sink, and shower(often appear simultaneously). And, the deterministic decoder may output them all together, and make mistakes. Admittedly, we also observe that deterministic sometimes perform better than temperature = 0.4, such as the buddy holly bus picture. However, the difference in bleu1 score is only reflected by predicting the word "bus", which is possibly due to temperature = 0.4 giving a bad choice because of luck. This scenario hardly occurs. Plus, considering the better performance of bleu1 score when temperature = 0.4 (71 to 66 bleu1 score), we believe deterministic predictions are worse.

8.3 Vanilla RNN

The result of RNN, 0.645 of BLEU-1 score, leads it to be the worst model. The architecture of RNN suffers the problem of gradient exploding and gradient vanishing when training with BPTT. So the weights updating in the cells may not reach the ideal situation. Thus, Vanilla performance is not good. On the other hand, from the perspective of loss plots, whatever combo of parameters we attempted, slight or obvious overfitting during the training process will always be happening. This might be another reason that vanilla RNN becomes the worst model among the three, which could be caused by the adjustment of layers of the framework compared with the baseline. Furthermore, considering the intervening of deterministic over the training process, it is another key element that drags the performance of this model down, so we preferably use the ones without temperature deterministic. Then, considering the examples we selected, regarding the variation of temperature between the default of 0.1, 0.001, and 5, not many big gaps of BLEU scores are detected, which is possibly due to the fair luck we have during the training.

8.4 Arch2

For images in architecture 2, the deterministic sampler still falls into the case of combining words that often appear simultaneously in the real world, to its actual output. For instance, in the image with a swimming pool, the deterministic sampler predicts "a bench and a bench on a beach." However, "beach" does not appear in the picture. "Beach" is predicted due to its close connection to "bench". And, we also believe that architecture 2 does not perform as good as the baseline LSTM model, since we detected a "repeating word pattern" in many of our outputs. For instance, for the image with several couches, architecture 2 outputs "a room with a couch, a couch, a couch, a couch, a couch,". We believe this output is due to consistently introducing the image features to the LSTM inputs. When the image contains lots of couches, the image features may strongly incline to the output "couches" in different time periods. Therefore, the word "couches" is predicted again and again, which is definitely not a desired output. Overall, we believe that the baseline LSTM model is the best among these three models.

9 Team Contributions

Wenbo Hu: Participated in implementation of all networks. Coding of stochastic generating caption, lstm and vanilla baseline. Tuning hyperparameters and finalizing results.
Wenxiao Li: wrote the code for samplers, bleu scores, good/bad image generation, tuned hyperparameters for the baseline LSTM model. Implemented the learning and testing procedures of architecture 2. Wrote reports for baseline LSTM and architecture 2.
Huaning Liu: Co-implementing for baseline model as well as vanilla RNN model. Hypertuning for RNN model. Organize the latex report, and write abstract, introduction, related work, model, and output analysis for Vanilla RNN model.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge," *arXiv e-prints*, p. arXiv:1609.06647, Sep. 2016.