Image Captioning in Practice: Using PyTorch and Recurrent Neural Network

Linghang Kong
Halıcıoğlu Data Science Institute
University of California San Diego
San Diego, CA 92093
l3kong@ucsd.edu

Weiyue Li
Halıcıoğlu Data Science Institute
University of California San Diego
San Diego, CA 92093
wel019@ucsd.edu

Yi Li
Halıcıoğlu Data Science Institute
University of California San Diego
San Diego, CA 92093
yil115@ucsd.edu

Shuangmu Hu
Jacobs School of Engineering
University of California San Diego
shh036@ucsd.edu

Yibo Wei
Jacobs School of Engineering
University of California San Diego
y2wei@ucsd.edu

Abstract

In this programming assignment, we trained an algorithm to caption input images, which requires the algorithm to identify what objects are in the images and how to match corpus of text to these objects. We have used PyTorch for multiple Recurrent Neural Network (LSTM, Vanilla RNN, and a customed model (architecture2)) and generated captions for the images in our dataset. The dataset we have used is the well-known COCO Image Captioning Task. Due to lack in the GPU resources, we have only used a portion of the original dataset to train, validate, and test. We have used cross entropy loss, BLUE-1 and BLUE-4 score to evaluate the performance of our model. Finally we were able to achieve final test loss 1.397, BLEU-1 score of 66.7% and BLEU-4 score of 7.69% using deterministic sampling technique, BLEU-1 score of 65.8% and BLEU-4 score of 7.7% using stochastic sampling technique under temperature 0.1, with best baseline LSTM model with hidden size 1024 and embedding size 300; Test loss of 1.480, BLEU-1 score of 65.0% and BLEU-4 score of 8.9% using deterministic sampling technique, BLEU-1 score of 68.3% and BLEU-4 score of 8.9% using stochastic sampling technique under temperature 0.001, with best Vanilla RNN model with hidden size 512 and embedding size 512; Test loss of 1.408, BLEU-1 score of 67.7% and BLEU-4 score of 8.9% using deterministic sampling technique, BLEU-1 score of 67.7% and BLEU-4 score of 8.9% using stochastic sampling technique under temperature 0.001, with best Architecture 2 Model with hidden size 1024 and embedding size 512. We can reach the conclusion that hidden size can have significant impact on model's performance, while embedding size does not have as much of an

effect; LSTM and Architecture 2, the models with more control-ability because of the gate mechanism, has better performance than Vanilla RNN; Passing in images at each time step does not help with performance when using cross entropy loss as evaluation metric, but such positive effect can be observed when BLEU score, a more appropriate evaluation metric for text interpretation, is introduced.

1 Introduction

With the development of the field of Computer Vision, computers' ability to process images are not limited to classify the images. Tasks such as detecting object in an image, or segment objects in image has been developing over the years. In this assignment, our goal is to generate captions for the images in the dataset. The captions have to be as descriptive of the images as possible. There might be a lot of ways of describing what is in the image, as there may be many different objects that can be detected in the images. The idea is to first use a pre-trained ResNet50 CNN as an encoder for the image, and then pass in text data as sequential data using recurrent neural network as decoder, thus generating captions for the input image. We used different alternatives for RNN, including LSTM, Vanilla RNN, and a customized model named Architecture 2. We used COCO Image Captioning Task dataset for training and testing. Only a portion of the whole dataset, which contains 82k images with 410k captions as training samples and 3k images with 15k captions as testing samples, will be used. We will evaluate the quality of generated captions using BLEU-1 and BLEU-4 scores, which are metrics commonly used in neural machine translation models such as BART or mBART. After fine-tuning our models, we were able to achieve final test loss 1.397, BLEU-1 score of 66.7% and BLEU-4 score of 7.69% using deterministic sampling technique, BLEU-1 score of 65.8% and BLEU-4 score of 7.7% using stochastic sampling technique under temperature 0.1, with best baseline LSTM model with hidden size 1024 and embedding size 300; Test loss of 1.480, BLEU-1 score of 65.0% and BLEU-4 score of 8.9% using deterministic sampling technique, BLEU-1 score of 68.3% and BLEU-4 score of 8.9% using stochastic sampling technique under temperature 0.001, with best Vanilla RNN model with hidden size 512 and embedding size 512; Test loss of 1.408, BLEU-1 score of 67.7% and BLEU-4 score of 8.9% using deterministic sampling technique, BLEU-1 score of 67.7% and BLEU-4 score of 8.9% using stochastic sampling technique under temperature 0.001, with best Architecture 2 Model with hidden size 1024 and embedding size 512. We can reach the conclusion that hidden size can have significant impact on model's performance, while embedding size does not have as much of an effect; LSTM and Architecture 2, the models with more control-ability because of the gate mechanism, has better performance than Vanilla RNN; Passing in images at each time step does not help with performance when using cross entropy loss as evaluation metric, but such positive effect can be observed when BLEU score, a more appropriate evaluation metric for text interpretation, is introduced.

2 Background/Related Work

To better fine-tune our model, we searched for related works on the similar tasks. A paper called *Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge* by Vinyals et.al[1]. in 2015 introduced a model-building and fine tuning similar to our assignment. In this paper, a model similar to our default LSTM model, called Neural Image Caption model (NIC) is introduced. In this paper, it is mentioned that "fine tuning the image model must be carried after the LSTM parameters have settled on a good language model", which has been good advice to our training process.

We also found out that Architecture 2 has some mechanisms that are used in attention-based image captioning models. Attention is a mechanism that allows image encoder to focus on different parts of the image, thus taking image spatial aspects into consideration. Such method was first used in Neural Machine Translation. *Sequence to sequence learning with neural networks* by Sutskever et. al[2]. introduced how this type of model works. Compared to our Architecture 2, apparently we do not have attention mechanism in Architecture 2, but images are processed at each time step in LSTM similar to the attention based model.

A new evaluating metrics, BLEU score, is introduced in this assignment. With a range of (0,1), it measures the similarity of generated captions to the actual captions. Geometric mean of the test corpus' modified precision scores is taken and multiplied by an exponential brevity penalty

factor. This score is originally used in Neural Machine Translation tasks, and was first introduced in Papineni et. al[3]. *BLEU: a Method for Automatic Evaluation of Machine Translation*.

3 Methods

3.1 Model Architecture

LSTM

The architecture of LSTM has two parts: the encoder and the decoder (shown in Figure 1). In the encoder part, we used a pre-trained convolutional network ResNet50 to extract features, and then add a fully connected layer and adjust the dimensions to the input of LSTM unit, which is the same as the word embedding dimensions. The weight of convolutional network is fixed. In the decoder part, we used a two layer LSTM with customized hidden size, with an embedding for each input word or image features from CNN. In the training, we used a method called Teacher-Forcing, which we utilize the target output instead of network output as the input for the next LSTM unit.

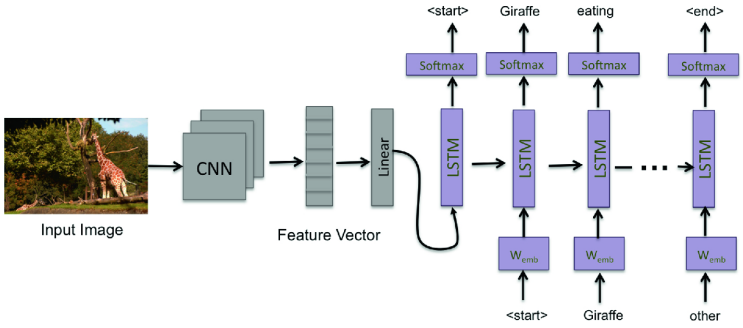


Figure 1: Encoder-Decoder architecture for image captioning (image credit: Deep Neural Network Based Image Captioning)

The default configuration of our baseline LSTM model is shown in the Table 1

Table 1: Default configuration of baseline LSTM model

CONFIGURATION	DEFAULT VALUE
img_size	256
batch_size	64
num_workers	1
num_epochs	10
learning_rate	5e-4
early_stop	0
hidden_size	512
embedding_size	300
model_type	”LSTM”
keep_image	false
max_length	20
deterministic	false
temperature	0.1
center_crop	0
horizontal_flip	true
rotation	10
normalize	true

Vanilla RNN

The vanilla RNN has basically the same overall architecture as the LSTM. It has an encoder that utilize the CNN to extract the features in the image, and connect the decoder with a fully connected embedding layers. The vanilla also has the two layers with customized hidden size and embedding size. The only difference is that it is a standard unit. It doesn't have any memory cells or gate, but takes simple input from the word embedding layers and the hidden layers, and output to the predicted word and the next hidden layer.

The default configuration of our Vanilla RNN model is shown in the Table 2

Table 2: Default configuration of Vanilla RNN model

CONFIGURATION	DEFAULT VALUE
img_size	256
batch_size	64
num_workers	1
num_epochs	10
learning_rate	5e-4
early_stop	0
hidden_size	512
embedding_size	300
model_type	"RNN"
keep_image	false
max_length	20
deterministic	false
temperature	0.1
center_crop	0
horizontal_flip	true
rotation	10
normalize	true

Architecture 2

Architecture 2 (shown in Figure 2) is the model with the image encoding given at each time step. It has a similar structure to our baseline LSTM. But instead of feeding encoded image to the decoder in the first time step, Architecture 2 passes encoded image to the decoder at every time step. In every step, the input will be a concatenation of expected output of previous step and encoded image.

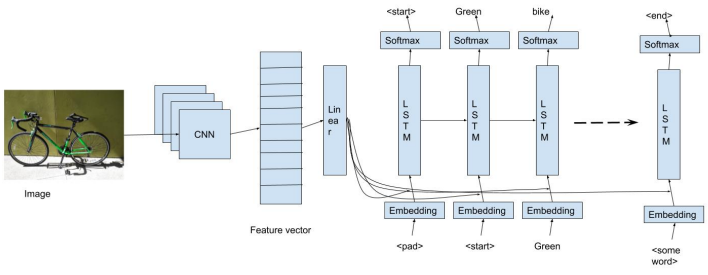


Figure 2: Architecture 2 for image captioning (image credit: CSE 151B course staff)

The default configuration of our Architecture 2 model is shown in the Table 3

Table 3: Default configuration of Architecture 2 model

CONFIGURATION	DEFAULT VALUE
img_size	256
batch_size	64
num_workers	1
num_epochs	10
learning_rate	5e-4
early_stop	0
hidden_size	512
embedding_size	300
model_type	"ARCH2"
keep_image	false
max_length	20
deterministic	false
temperature	0.1
center_crop	0
horizontal_flip	true
rotation	10
normalize	true

3.2 Output Sampling

To sample the output and generate the captions, we let the network run on its own. That is, we input the image, predict the first word, and use the word to predict the next word until the end, which is different than "teacher forcing" that we used in the training procedure. To print the actual words, we have two approaches. One is the deterministic approach, where we take the maximum output, and another one is stochastic, where we randomly choose one word by their probability distribution. The probability distribution is obtained by the softmax function with temperature using equation $y^j = exp(o^j/\tau)/\sum_n exp(o^n/\tau)$. Different from our default LSTM and Vanilla RNN models, when sampling from Architecture 2, we need to initialize padding and concatentate images and texts before the sampling process begins.

3.3 Parameter Search Conduction

Due to the time restriction, we mainly tuned the hidden size and the embedding size. We choose three hidden size and three embedding size, and cross-combine these two hyperparameters, which is 3×3 combinations for each model in total. After we obtained our best result from changing hidden size and embedding size, we used the set of parameters from our best models to adjust learning rate. Besides that, we also randomly tried changing other parameters, like replacing the Adam Optimizer with SGD. However, these changes barely improves the performance of the model.

3.4 Word Embedding

Caption texts are first one-hot encoded and then converted to a low dimension embedding through a fully connected layer before putting into LSTM as a current feature. We took advantage of PyTorch's nn.Embedding to simplify this process. 'Pad' was manually added to the embedded input for Architecture 2, and the embedded caption was passed in with embedded images.

4 Results

4.1 Hyperparameters

After increasing hidden size from default value 512 to 1024 and maintaining the default value for embedding size, it can improves the performances for model LSTM. The validation loss for LSTM decreases from 1.41825 (using default settings) to 1.37344. After increasing embedding size from

default value 300 to 512 and maintaining the default value for hidden size, it improves the performances for model VANILLA. The validation loss for VANILLA decreases from 1.48365 (using default settings) to 1.47102. After increasing both hidden size from default value 512 to 1024 and embedding size from default value 300 to 512, it improves the performances for model ARCH2. The validation loss for ARCH2 decreases from 1.4224 (using default settings) to 1.38731. After finding the best combination of hidden size and embedding size, we decrease the learning rate for those combinations. However, decreasing the learning rate do not further improve the performances for all three models even taking more epochs. The corresponding data are shown in Table 4. Moreover, Table 5, 6, and 7 show the best models' configurations accordingly.

Table 4: Validation loss of different hidden sizes and models

HIDDEN SIZE	EMBEDDING SIZE	LEARNING RATE	NUMBER OF EPOCH	LSTM	VANILLA	ARCH2
512	256	$5e^{-4}$	10 (converge)	1.4186	1.4712	1.4480
256	300	$5e^{-4}$	10 (converge)	1.415091	1.49192	1.45093
512	300	$5e^{-4}$	10 (converge)	1.41825	1.48365	1.4224
1024	300	$5e^{-4}$	10 (converge)	**1.37344**	1.50572	1.39850
512	512	$5e^{-4}$	10 (converge)	1.40891	**1.47102**	1.4288
1024	512	$5e^{-4}$	10 (converge)	1.39113	1.48202	**1.38731**
1024	300	$1e^{-4}$	13 (converge)	1.41534	N/A	N/A
512	512	$1e^{-4}$	12 (converge)	N/A	1.49422	N/A
1024	512	$1e^{-4}$	14 (converge)	N/A	N/A	1.43513

Table 5: Best configuration of baseline LSTM model

CONFIGURATION	BEST VALUE
img_size	256
batch_size	64
num_workers	1
num_epochs	10
learning_rate	5e-4
early_stop	0
hidden_size	1024
embedding_size	300
model_type	"LSTM"
keep_image	false
max_length	20
deterministic	true
temperature	0.4
center_crop	0
horizontal_flip	true
rotation	10
normalize	true

Table 6: Best configuration of Vanilla RNN model

CONFIGURATION	BEST VALUE
img_size	256
batch_size	64
num_workers	1
num_epochs	10
learning_rate	5e-4
early_stop	0
hidden_size	512
embedding_size	512
model_type	”RNN”
keep_image	false
max_length	20
deterministic	true
temperature	0.4
center_crop	0
horizontal_flip	true
rotation	10
normalize	true

Table 7: Best configuration of Architecture 2 model

CONFIGURATION	BEST VALUE
img_size	256
batch_size	64
num_workers	1
num_epochs	10
learning_rate	5e-4
early_stop	0
hidden_size	1024
embedding_size	512
model_type	”ARCH2”
keep_image	false
max_length	20
deterministic	false
temperature	0.4
center_crop	0
horizontal_flip	true
rotation	10
normalize	true

4.2 Losses

When conducting our experiments, we have recorded the training & validation loss for each epoch. We will be reporting them in the sub-sub sections below accordingly.

LSTM

For the LSTM model, it performs the best when using the hidden size of 1024 and the embedding size of 300. The training loss was greater than the validation loss initially due to random chance. However, starting around the second epoch after some sort of training, the validation loss has become greater than the training loss. As we can see in Figure 3, the validation loss starts to converge after 4 epochs for the LSTM model. And, the model starts to over-fit after the 6th epoch.

Table 8 shows the training & validation loss for the baseline LSTM model.

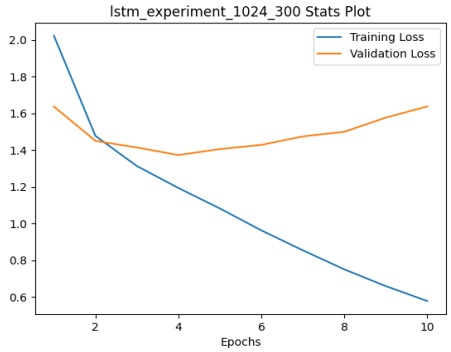


Figure 3: Training and validation loss for the baseline LSTM model

Table 8: Training and validation loss for the baseline LSTM model

EPOCH	TRAINING LOSS	VALIDATION LOSS
1	2.02377	1.63758
2	1.47790	1.45071
3	1.31336	1.41440
4	1.19379	1.37343
5	1.08208	1.40586
6	0.96257	1.42887
7	0.85388	1.47492
8	0.75006	1.50027
9	0.65874	1.57750
10	0.57738	1.63832

Vanilla RNN

For the vanilla RNN model, it performs the best when using hidden size of 512 and the embedding size of 512. As the LSTM model, the vanilla RNN's training loss is also greater than validation at the beginning, and the validation loss starts to converge after 6 epochs. The Figure 4 shows the loss curve, while Table 9 shows the loss data.

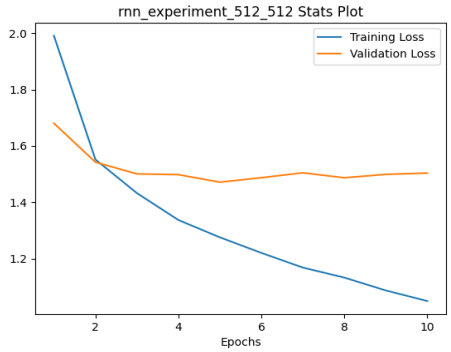


Figure 4: Training and validation loss for the Vanilla RNN model

Table 9: Training and validation loss for the Vanilla RNN model

EPOCH	TRAINING LOSS	VALIDATION LOSS
1	1.99098	1.68020
2	1.55165	1.54216
3	1.43219	1.50041
4	1.33669	1.49779
5	1.27518	1.47102
6	1.21983	1.48694
7	1.16777	1.50432
8	1.13219	1.48664
9	1.08654	1.49880
10	1.04894	1.50332

Architecture 2

For the Architecture2 model, it performs the best using hidden size of 1024 and embedding size of 300. Same as previous two models, the loss curves performs as our expected, which means the training loss was higher than validation loss at the beginning, and it starts to converges after 4 epochs. The Figure 5 and Table 10 shows the result.

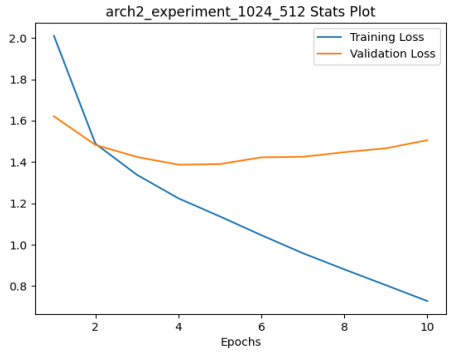


Figure 5: Training and validation loss for the Architecture 2 model

Table 10: Training and validation loss for the Archetecture2 model

EPOCH	TRAINING LOSS	VALIDATION LOSS
1	2.05912	1.65794
2	1.52066	1.50174
3	1.36654	1.44346
4	1.25093	1.42244
5	1.15657	1.40307
6	1.07116	1.39850
7	0.99010	1.42925
8	0.90749	1.43109
9	0.83133	1.49274
10	0.75088	1.50318

All Training and Validation Loss for Best Models

The training and validation loss for all best models (LSTM, Vanilla RNN, and Architecture 2) is shown in Figure 6. The general trend of all best models is starting to over-fit to the training set after 4-5th epochs. Out of the best models, the baseline LSTM has the best validation loss after fine-tuning, followed by Architecture 2 and Vanilla RNN accordingly.

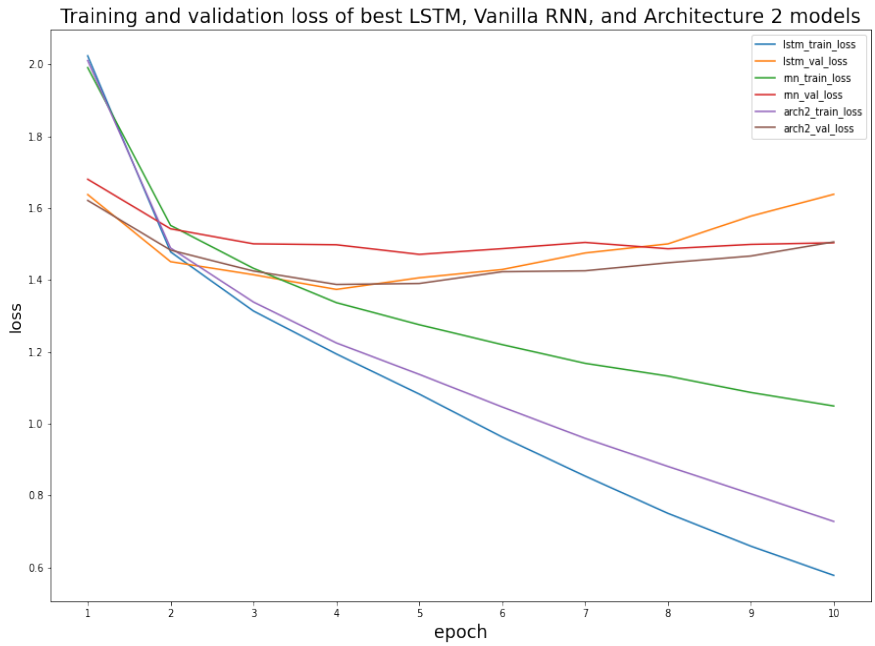


Figure 6: Training and validation loss for all best models

Test Loss for Best Models

Table 11 shows the cross-entropy test loss for all of our best models.

As shown in the table, LSTM has the best test loss of 1.397, followed by Architecture 2 1.408, and lastly Vanilla RNN with loss of 1.480. The interpretation can be seen in the discussion section.

Table 11: Cross entropy test loss for all best models

MODEL	TEST LOSS
LSTM	1.396781526981516
Vanilla RNN	1.4804498634439833
Architecture 2	1.4075451594717958

4.3 BLEU Scores

Bilingual Evaluation Understudy score, or BLEU score for short, is a metric that we have utilized for evaluating our model-generated captions. The BLEU score metric is widely adopted because it is language independent, inexpensive and it has high correlation with human evaluated scores. By

definition, it ranges from 0 to 1 with 0 meaning it is completely different from the referred captions and 1 meaning it is completely matching the referred captions. According to the original BLEU score paper[3], even human beings cannot achieve a BLEU score of 1 for a large number of images.

When conducting our experiments, we have recorded the BLEU-1 and BLEU-4 scores for each trained model. We will be reporting them in the sub-sub sections below accordingly.

LSTM

For our baseline LSTM model, we were able to achieve 0.658 of BLEU-1 score and 0.077 of BLEU-4 score in our best model when using stochastic generating techniques. Some of the key observations is that when decreasing temperature, the BLEU-1 score were able to increase. This is why our BLEU-1 score were able to reach 0.667 with the same setting when using deterministic generating techniques because as temperature approaches 0, the distribution converges deterministically.

Table 12 below shows the BLEU-1 and BLEU-4 scores of the baseline LSTM model in default setting with different temperature in two different techniques.

Table 12: BLEU scores of best baseline LSTM model with different temperatures

TECHNIQUE	TEMPERATURE	BLEU-1 SCORE	BLEU-4 SCORE
Stochastic	5	0.009217888255893418	0.0009964008769568998
Stochastic	1	0.4712727196868304	0.025361589946011683
Stochastic	0.4	0.6350861666176466	0.06576195962668056
Stochastic	0.1	0.6579426598446451	0.07667242393766965
Stochastic	0.001	0.6573381160778891	0.07725509405253467
Deterministic	-	0.6669803889363306	0.07689079015736867

Vanilla RNN

For our Vanilla RNN model, we were able to achieve 0.650 of BLEU-1 score and 0.074 of BLEU-4 score in our best model when using stochastic generating techniques. Some of the key observations is that when decreasing temperature, the BLEU-1 score were able to increase. This is why our BLEU-1 score were able to reach a similar 0.650 with the same setting when using deterministic generating techniques because as temperature approaches 0, the distribution converges deterministically. Another finding is that the BLEU scores of this Vanilla RNN model is generally lower than the BLEU score of baseline LSTM model. This is because LSTM has reduced the vanishing gradient problem by adding an memory cell, which has the effect of having higher BLEU scores of Vanilla RNN.

Table 13 below shows the BLEU-1 and BLEU-4 scores of the Vanilla RNN model in default setting with different temperature in two different techniques.

Table 13: BLEU scores of best Vanilla RNN model with different temperatures

TECHNIQUE	TEMPERATURE	BLEU-1 SCORE	BLEU-4 SCORE
Stochastic	5	0.009824861888107264	0.0010539837051789444
Stochastic	1	0.4583802878525096	0.024369451851921554
Stochastic	0.4	0.6301167654169872	0.06362408887193974
Stochastic	0.1	0.649773666425041	0.07333452087431787
Stochastic	0.001	0.6502439110303553	0.07409273324906046
Deterministic	-	0.6503196357355883	0.07414966468197391

Architecture2

For our Architecture 2 model, we were able to achieve 0.677 of BLEU-1 score and 0.087 of BLEU-4 score in our best model when using stochastic generating techniques. Some of the key observations is that when decreasing temperature, the BLEU-1 score were able to increase. This is why our BLEU-1 score were able to reach 0.677 with the same setting when using deterministic generating techniques because as temperature approaches 0, the distribution converges deterministically. The reason why architecture 2 has better performances than the rest models is that architecture 2 will concatenate the image with the outputs from the last time-step for each time-step. Therefore, architecture 2 will be able to learn more details from the concatenate inputs during training. **However,** we are able to obtain the **0.688** of BLEU-1 score and **0.089** of BLEU-4 score in our Architecture 2 model with 512 hidden size and 250 embedding size. As a result, we suppose that the validation loss for our models may not correlated with the BLEU scores.

Table 14 below shows the BLEU-1 and BLEU-4 scores of the Architecture 2 model in default setting with different temperature in two different techniques.

Table 14: BLEU scores of best Architecture 2 model with different temperatures

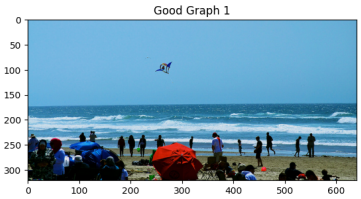
TECHNIQUE	TEMPERATURE	BLEU-1 SCORE	BLEU-4 SCORE
Stochastic	5	0.009929644226494127	0.000997563622116251
Stochastic	1	0.4770911540036202	0.0244665095052084
Stochastic	0.4	0.6774908680063205	0.08694107041395088
Stochastic	0.1	0.6764035124704819	0.08643270965482822
Stochastic	0.001	0.6773504418116775	0.08686583333802806
Deterministic	-	0.6774908680063206	0.08694107041395087

4.4 Examples

To further investigate our models' performances, we have visualized some of the sample captions for all three of our models. The images are displayed in the below sub-sub sections.

According to the graphs below, the lower the temperature, the more understandable and accurate the predicted texts. Such trend can be clearly observed in Figure 20. When temperature is large, the predicted text is just random words combined together. When temperature was reduced to 0.4, the description is more precise, with correct objects, gestures, and relative positions. However, the model did not use the correct quantity. When temperature was decreased to 0.001, the correct quantity was chosen. In this circumstance, a model with a temperature of 0.001 performs as accurate as deterministic sampling. This phenomenon matches our test result where deterministic sampling has the best performance in general. Similar results can be observed in the bad graphs. For example, in Figure 22, even though the BLEU score is low, we can still see how more accurate texts are predicted as we lower the temperature. At a temperature of 5, the model predicted almost random texts; When temperature drops to 0.4 and 0.001, the texts made sense, though it was not an accurate description.

LSTM



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a group of people flying kites on a beach ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a group of people standing on top of a sandy
beach ."

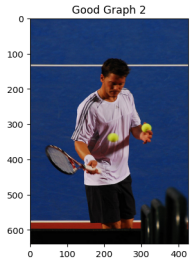
When temperature is: 5
When mode is: stochastic
Predicted text will be: "handheld directional fried countryside struggle tours louvered stealth way bolts
ways partially-eaten rubber object paper eighteen bicyclist hogs longhaired marshmallow"

When mode is: deterministic
Predicted text will be: "a group of people flying kites on a beach ."

Actual text: (1): a group of people standing on top of a beach near the ocean .
(2): a group of people at the beach flying kites
(3): people relaxing on the beach watching a kite .
(4): the people at the beach are enjoying watching the kite .
(5): several people on the beach looking at something in the sky .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.14286

Figure 7: LSTM sample good prediction 1 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man holding a tennis racquet on a tennis court ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a man holding a tennis racket in the air ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "near mannequins climbs apart ethnic conversing servings screen waterskiing knick
sunday . crowd stand-up as pew instructional pone decked kits"

When mode is: deterministic
Predicted text will be: "a man holding a tennis racquet on a tennis court ."

Actual text: (1): a man holding a tennis racquet and two tennis balls .
(2): a man looking at the tennis ball his is hitting .
(3): a male tennis player serving the green ball .
(4): man holding tennis racket with two balls in the air .
(5): a man on a tennis court holding a racket and two tennis balls .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.5

Figure 8: LSTM sample good prediction 2 at temperature = 0.4

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man riding a motorcycle with a helmet on ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a man riding a motorcycle with a dog on his
bike"

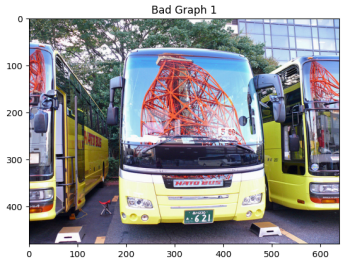
When temperature is: 5
When mode is: stochastic
Predicted text will be: "arguing organization delightful conversations vending horned apparently load heavenly lined
perhaps full sightseeing ottoman branches dj champagnes sand stalk gracefully"

When mode is: deterministic
Predicted text will be: "a man riding a motorcycle with a helmet on ."

Actual text: (1): a person riding a motorcycle on a road
(2): a person on a motorbike has a helmet on .
(3): a man with a leather jacket sitting on a motorcycle in a street .
(4): a male in a black jacket is on his motorcycle and some bushes and grass behind him
(5): a man making a turn on a motorcycle and looking back .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.14286

Figure 9: LSTM sample good prediction 3 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a double decker bus is driving down the street ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a red double decker bus driving down a street ."

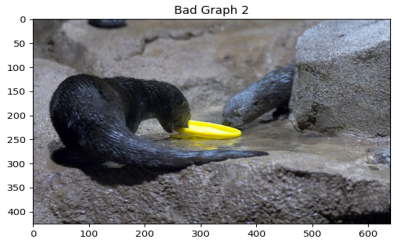
When temperature is: 5
When mode is: stochastic
Predicted text will be: "carriage overview appliance focus taht cables hangings shopping windows headscarf
hurtle cars staged getting confused lite screened creme sample dots"

When mode is: deterministic
Predicted text will be: "a double decker bus is driving down the street ."

Actual text: (1): three yellow coach buses parked in a line .
(2): a bus with a reflection of metal work on the windshield .
(3): parked buses have a tower reflected in the window .
(4): a group of busses parked in a parking lot .
(5): the reflection of a large metal structure in the windshield of a bus

When temperature is 0.4, bleu1 score is: 0.4
When temperature is 0.4, bleu4 score is: 0.01429

Figure 10: LSTM sample bad prediction 1 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a bird sitting on a branch of a rock ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a polar bear sitting on a rock , in the
water ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "enjoying llittle pops tweed kitesurfing grind ahead winding ground cork
wonderful rippling liner adventure before back can identically gloomy stew"

When mode is: deterministic
Predicted text will be: "a bird sitting on a branch of a rock ."

Actual text: (1): an animal biting a yellow frisbee next to another man .
(2): otters investigating a frisbee thrown into their naturalistic enclosure .
(3): two dark colored animals with a yellow plastic disc .
(4): two otters that are playing with a frisbee .
(5): two small otters playing with a yellow frisbee .

When temperature is 0.4, bleu1 score is: 0.2
When temperature is 0.4, bleu4 score is: 0.01429

Figure 11: LSTM sample bad prediction 2 at temperature = 0.4

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man riding a bike down a street ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a group of people riding motorcycles down a street ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "serve rooom parked wildnerness tails eagerly barns beds designating corner
distributing night sale flashing minimal coupons handles standalone mixed plan"

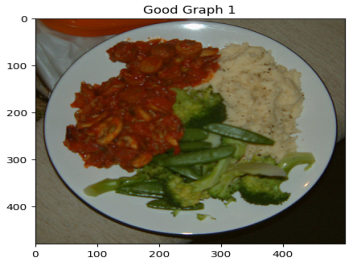
When mode is: deterministic
Predicted text will be: "a man riding a bike down a street ."

Actual text: (1): this is a group of people standing near a river
(2): a group of people with bikes posing for a photo
(3): seven people on a biking trip in front of a large city .
(4): group of bikers posing for a picture
(5): a bunch of people posing with some bikes .

When temperature is 0.4, bleu1 score is: 0.33333
When temperature is 0.4, bleu4 score is: 0.01667

Figure 12: LSTM sample bad prediction 3 at temperature = 0.4

Vanilla RNN



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a plate of food with broccoli and meat ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a plate of food that includes broccoli and broccoli ."

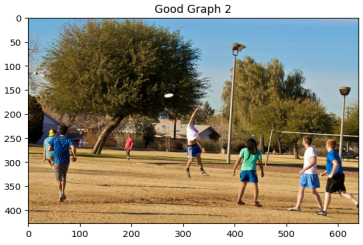
When temperature is: 5
When mode is: stochastic
Predicted text will be: "present called crumbs ext panel fuel falafel hugged rests dek fillet grove ' vitamins craning . knife frowning serving octopus"

When mode is: deterministic
Predicted text will be: "a plate of food with broccoli and meat ."

Actual text: (1): a plate holds dinner including green beans and potatoes .
(2): meat and vegetables are on a white plate .
(3): a white plate covered in sauce covered food rice and broccoli .
(4): a plate of food with pasta , mashed potatoes and broccoli .
(5): a plate filled with three different types of foods .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.33333

Figure 13: Vanilla RNN sample good prediction 1 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a group of people playing frisbee in a park ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a group of people playing frisbee in a park ."

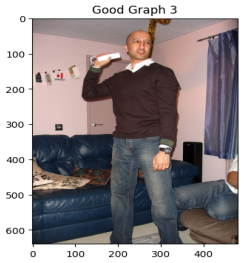
When temperature is: 5
When mode is: stochastic
Predicted text will be: "articles boxcar tossing rich trophies accompanied lama pastor outskirts english major murdering taxi isnide sport parking choco wndow leafy wide"

When mode is: deterministic
Predicted text will be: "a group of people playing frisbee in a park ."

Actual text: (1): many people are playing a game and having fun .
(2): a group of people playing a game of frisbee .
(3): kids playing frisbee in a park on a bright day
(4): teens out in the park playing a game of frisbee
(5): a group of people playing frisbee in a field

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.85714

Figure 14: Vanilla RNN sample good prediction 2 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man standing in a living room playing wii"

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a man sitting on a couch with a wii controller ."

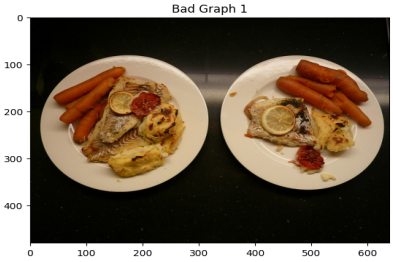
When temperature is: 5
When mode is: stochastic
Predicted text will be: "surer bureau sixteen sea accessory dim brown hold walkout crawling crouch social suitcases majestic interactive herbs genius ultimate train temporary"

When mode is: deterministic
Predicted text will be: "a man standing in a living room playing wii"

Actual text: (1): a man is playing a wii video game .
(2): a man standing in a living room holding a wii controller .
(3): the man is using his remote to enjoy the game .
(4): a male in a black shirt i splaying a video game
(5): a man in brown sweater playing a game with nintendo wii controller .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.66667

Figure 15: Vanilla RNN sample good prediction 3 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a plate with a sandwich , pickle and a pickle ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a white plate topped with a sandwich and a bowl of salad ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "brach thinly unappealing freeways j gooey referee wheat gloves broth seen dougnuts .... confronting slipped insignia boarding hamburger writting reds"

When mode is: deterministic
Predicted text will be: "a plate with a sandwich , pickle and a pickle ."

Actual text: (1): two white plates topped with meals sitting on a table .
(2): two white plates filled with steamed carrots and grilled fish .
(3): two white plates holds cooked fish and carrots .
(4): twp plates of food are set next to each other .
(5): a close up of two plates contaning fish and carrots

When temperature is 0.4, bleu1 score is: 0.36364
When temperature is 0.4, bleu4 score is: 0.0125

Figure 16: Vanilla RNN sample bad prediction 1 at temperature = 0.4

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a bag with a cell phone and a bag on
it ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a bag and a bag on a sidewalk ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "fiskars message firsbee uprooted oral hunters tangled whole strewn stacks
fender aluminium frames table ground ladybug under cosplaying microphones macbook"

When mode is: deterministic
Predicted text will be: "a bag with a cell phone and a bag on
it ."

Actual text: (1): a couple bags of luggage that are on the ground .
(2): four different eras of luggage are tossed along a rail .
(3): four different types of carrying cases in various colors .
(4): four pieces of luggage sit on the ground .
(5): a pile of luggage sitting on the ground at an airport .

When temperature is 0.4, bleu1 score is: 0.25
When temperature is 0.4, bleu4 score is: 0.01111

Figure 17: Vanilla RNN sample bad prediction 2 at temperature = 0.4

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man is working on a laptop in a kitchen ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a man standing in front of a store filled with
lots of luggage ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "facades acknowledging environment neutral group addresses papers unfinished frame as
orderly gondala tangled swatting opponents snake carried ruffle just or"

When mode is: deterministic
Predicted text will be: "a man is working on a laptop in a kitchen ."

Actual text: (1): people at the baggage claim area of an airport .
(2): three people standing before airport counters below airport signs .
(3): people standing at counters of booths being served
(4): the baggage delivery section of an air port .
(5): patrons are going to the shops of an airport .

When temperature is 0.4, bleu1 score is: 0.09091
When temperature is 0.4, bleu4 score is: 0.0125

Figure 18: Vanilla RNN sample bad prediction 3 at temperature = 0.4

Architecture2

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a group of people riding horses along the beach ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a group of people riding horses along the beach ."

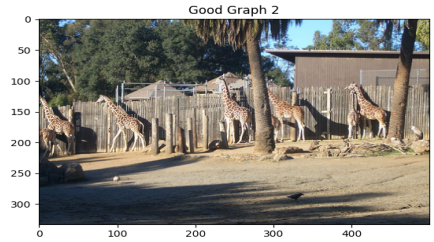
When temperature is: 5
When mode is: stochastic
Predicted text will be: "aged skiiers turmac carriages tips tracks instruct mill stairways temporary flaps related base decortive n itching up scrubs creating tolls"

When mode is: deterministic
Predicted text will be: "a group of people riding horses along the beach ."

Actual text: (1): a group of people horse back riding on a sandy beach .
(2): group of people riding horses on the side of the beach .
(3): a group of people riding on horses on a beach
(4): four people riding horses next to the water on a beach
(5): a group of people on horses walk along the beach .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.57143

Figure 19: Architecture 2 sample good prediction 1 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a group of giraffes standing next to a fence ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "two giraffes stand in a zoo enclosure near a fence ."

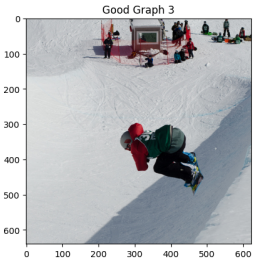
When temperature is: 5
When mode is: stochastic
Predicted text will be: "formation alot hang stands circular harsh pouting tied head . medicine poll patricks frisk soccor barrel lie planting baskets directing"

When mode is: deterministic
Predicted text will be: "a group of giraffes standing next to a fence ."

Actual text: (1): a group of giraffes next to a fence .
(2): a few zebras are standing in front of the fence .
(3): many giraffes walk along a wooden fence in a small lot with sand and palm trees .
(4): a group of giraffes that are walking near a fence .
(5): there are many giraffes that are standing by a fence

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 0.42857

Figure 20: Architecture 2 sample good prediction 2 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man riding a snowboard down a snow covered slope ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a man riding a snowboard down a snow covered slope ."

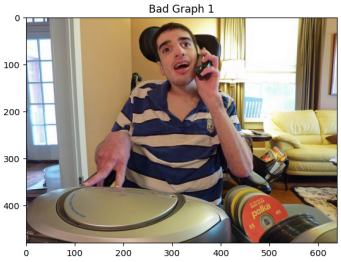
When temperature is: 5
When mode is: stochastic
Predicted text will be: "garlic demonstrate relish top performance protective sombrero filthy ramshackle logo chalice snowing zombie lady listening upside overlooks stunts rapped ultimate"

When mode is: deterministic
Predicted text will be: "a man riding a snowboard down a snow covered slope ."

Actual text: (1): a man riding a snowboard down a snow covered slope .
(2): an image of a person on skiis going the slope
(3): a snow boarder doing a small jump by the side of a hill
(4): a person jumping a snow board in the air
(5): a snowboarder does a trick in the air .

When temperature is 0.4, bleu1 score is: 1.0
When temperature is 0.4, bleu4 score is: 1.0

Figure 21: Architecture 2 sample good prediction 3 at temperature = 0.4



When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a woman in a kitchen preparing food"

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a woman in a kitchen with a piece of food
on it ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "blocks cross-country scooping needle admires friends wrapper at confusedly confetti wrigley pet windsurfing conversations basketball veggie easel untouched live titled"

When mode is: deterministic
Predicted text will be: "a woman in a kitchen preparing food"

Actual text: (1): a person in a wheelchair plays music on a cd player while talking on a wireless telephone .
(2): a man in a wheelchair holding a phone to his ear , in front of a cd player and cd 's on the side .
(3): a man sitting at a table in a wheelchair while on a phone .
(4): a man sitting in front of a ghetto blaster while talking on a phone .
(5): a person sitting in a hair talking on a cell phone

When temperature is 0.4, bleu1 score is: 0.24202
When temperature is 0.4, bleu4 score is: 0.01412

Figure 22: Architecture 2 sample bad prediction 1 at temperature = 0.4

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a man is riding a horse down a street ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a person is holding a bat in a park ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "mechanical battered steele protest tennis gallops eagerly leans moldings appliances ms rin brass crouch living donations caution same individual airline"

When mode is: deterministic
Predicted text will be: "a man is riding a horse down a street ."

Actual text: (1): a group of people standing around at a rodeo .
(2): a rodeo area with men in cowboy hats , rodeo clowns and an audience in the stands .
(3): a contestant has caught a sheep during a rodeo event .
(4): we are looking through the bars of a fence into a rodeo arena .
(5): a young child mutton busting at a rodeo event

When temperature is 0.4, bleu1 score is: 0.4
When temperature is 0.4, bleu4 score is: 0.01429

Figure 23: Architecture 2 sample bad prediction 2 at temperature = 0.4

When temperature is: 0.001
When mode is: stochastic
Predicted text will be: "a woman is holding a pink umbrella and a woman ."

When temperature is: 0.4
When mode is: stochastic
Predicted text will be: "a woman and a woman sitting on a bench ."

When temperature is: 5
When mode is: stochastic
Predicted text will be: "democracy shoeless donut draining previous stunt luggage themed freighter span gloves coming ruins wiffle hoses others stand severed wishes wahlberg"

When mode is: deterministic
Predicted text will be: "a woman is holding a pink umbrella and a woman ."

Actual text: (1): a collaboration of people in different pictures doing things
(2): a series of images of young men painting and holding kites .
(3): various children and adults are making their own kites .
(4): a variety of colorful pictures with people doing various activities .
(5): a few people working with colored fabrics in different ways .

When temperature is 0.4, bleu1 score is: 0.36364
When temperature is 0.4, bleu4 score is: 0.0125

Figure 24: Architecture 2 sample bad prediction 3 at temperature = 0.4

5 Discussion

5.1 Model Performance

According to table 4, the LSTM model has the best performance with hidden size 1024 and embedding size 300 under learning rate $5e^{-4}$, Vanilla RNN has best performance with hidden size 512 and embedding size 512 under learning rate $5e^{-4}$, and Architecture 2 has best performance with hidden size 1024 and embedding size 512 $5e^{-4}$. If we compare model performance with different embedding size but fixed hidden size, we observed that the loss does not have significant variation due to change in embedding size. When hidden size increases, the model's performance can be greatly boosted. Due to embedding size's subtle influence to the model, the small difference between best performance of LSTM and Architecture 2 can be ignored when compared to the improvement the parameter have on the models. Even though Vanilla RNN performed the best with hidden size 512 and embedding size 512, which seems to counter our observation, the best performance the RNN model has cannot even match the performance of LSTM and Architecture 2 with default parameters. Thus we should not evaluate their performance equally in our conclusion. We can conclude that the best models are the way they are because they have larger hidden sizes. LSTM and Architecture 2 have better performance than RNN because they have gate control units. On the other hand, Architecture 2 does not have a lower validation loss or test loss compared to LSTM, it is reasonable to infer that passing in encoded image with captions at each time step does not improve model's performance when using cross entropy loss as an evaluation metrics. However, when evaluating model's performance using BLEU score on the test set, Architecture 2 has much more advantage than LSTM and Vanilla RNN. Since data like sentences cannot be judged by quantified values solely due to variations in grammar and diction, BLEU score is definitely a better way to evaluate how model performed in describing the images given, and Architecture 2, the model that passes in encoded images together with captions at each time step, indeed produces better result than LSTM and Vanilla RNN.

5.2 Generated Caption Results

Based on our test results and example visualizations, we do not believe that deterministic approach does not work well, at least in our cases. Besides the fact that deterministic approach has better BLEU scores than most of the stochastic approach models at different temperatures, as the temperature value decreases, a stochastic approach becomes more deterministic, and those stochastic sampling with lower temperatures has better performance than higher temperatures in our results. This raising performance with decreasing temperature value confirms our conclusion that we do not believe that deterministic approach does not work well in our cases. We have the assumption that, theoretically, the larger the temperature value, the more uniform the distribution of probability is, thus the more random the sampling process. So that in this case, stochastic sampling only provides more possible outcomes in which better generated captions may appear, but it does not necessarily guarantee a better overall result.

According to the visualized examples, it is apparent to conclude that the images' quality also has significant influence on model's performance. The "good graphs" usually have easily-separable pixels and simple position-wise or action-wise relationships. The objects can be clearly seen by the pretrained models we used and have a certain degree of contrast with the background. "Bad graphs", on the other hand, have much more varying pixels on average that make objects more difficult to be precisely detected. In Figure 10, a tower like object is reflected on the bus' windshield, adding more irrelevant pixels for the model and increases detection difficulty; So that the model may relate the tower building to words that has more association with bus such as "street", thus producing texts like "driving down a street" while the bus is actually parked. In Figure 11, besides the fact that the animal is not an easily-recognizable one, the Frisbee may have affected the model's decision, and the model mistakenly believe that the Frisbee is part of the animal, thus generating word "bird" since bird has such appearance. In other circumstances, a picture not only has complicated pixel components, but also has no main focus. Because of this lack of focus, the actual text concentrates on many different aspects of the image instead of having a common target. Just as it is shown in Figure 17 and Figure 18, the actual texts have many different ways of interpreting the images. In such cases, the predicted texts can hardly be given a good BLEU score.

5.3 Potential Future Improvement

In this assignment, we mainly implemented the models and fine-tuned the model per homework instruction. However, due to lack of GPU resources and limited time, we do not have the chance to try all the hyperparameters we want. Based on the information found on the internet, when modeling on the COCO dataset, we can also consider change the encoder to something other than ResNet (Vinyals et.al.). Large text-based language processing models such as BERT or GPT may also be combined with the image encoder to produce unexpected results.

6 Team Contribution

1. Linghang Kong: Initialized basic models for training and validation, assisted the other teammates in implementing test, sampling, visualization, as well as debugging the code. In charge of Abstract, Introduction, Related work, and part of Model Architecture and Discussion section of the report.
2. Weiyue Li: Helped implementing `model_factory.py` and `experiment.py`. In charge of implementing deterministic, train, val, and visualization methods, documenting the source code, and writing the `README.md`. Assisting teammates with report.
3. Yi Li: Implemented `model_factory.py` and `experiment.py`. In charge of implementing test, stochastic, and visualization methods and documenting the source code. Assisting teammates with report.
4. Shuangmu Hu: Fully participated in coding, help implementing `experiment.py` and caption generation. In charge of training and fine tuning. Assisting teammates with report.
5. Yibo Wei: Participated in implementing data augmentation in `dataset_factory.py`, stochastic generation in `model_factory.py`, test function in `experiment.py`, debugging Architecture 2, and writing the report.

Acknowledgments

We appreciate the help from course Piazza and TA&Tutor office hours, as well as the knowledge gained from Professor Garrison W. Cottrell's lectures.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. volume 39, page 652–663. Institute of Electrical and Electronics Engineers (IEEE), Apr 2017.

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.