Image Captioning with LSTM

Noah Inada
UC San Diego
CSE 151B
ninada@ucsd.edu

Sheung Ho
UC San Diego
CSE 151B
s4ho@ucsd.edu

Kaiwen Tan
UC San Diego
CSE 151B
kat066@ucsd.edu

Abstract

In this project, we developed an LSTM network that is able to caption images and compared it with other recurrent neural networks. We found that an LSTM performs better than an RNN, and that deterministically generated captions is usually better than using a stochastic approach.

1 Introduction

We built a model that can generate captions for images using an encoder/decoder architecture, encoding the images into vectors of feature values and passing those vectors through an LSTM network. A pre-trained resnet50 network was used for encoding the images. We trained on one-fifth of the COCO (Common Objects in Context) dataset, and so used 82k images from training and 3k images for testing. Each image has five human-generated captions. We compared our LSTM network with a vanilla RNN network and with LSTM/RNN networks with differing hyperparameters.

2 Related Work

CSE 151B discussion notebook

pytorch.org

https://towardsdatascience.com/automatic-image-captioning-with-cnn-rnn-aae3cd442d83

3 Architecture for baseline(LSTM) vanilla RNN model.

Our baseline(LSTM) model uses a ResNet50 pretrained model as an encoder. The last layer of the ResNet50 model is removed and replaced with a trainable linear model to map from the penultimate hidden layer dimension to the hidden LSTM dimension. Then, the encoded feature of the image is fed into a linear layer with an output size equal to a predetermined embedding size. All the words in each caption is fed into an embedding layer, and then the encoded image is concatenated with those embedded captions and fed into the LSTM layers. To generate captions, we use the image of the first timestamp to generate the first word, and then feed that word as input to generate the next word, and so on until a maximum caption length is read or until an ‘<end>’ token is generated.

For our fine-tuning, we tried using the default hyperparameters for our vanilla RNN and architecture 2 models.

Vanilla RNN model: the architecture of the RNN model is almost identical to the baseline model. except the LSTM layer is replaced with the RNN layer.

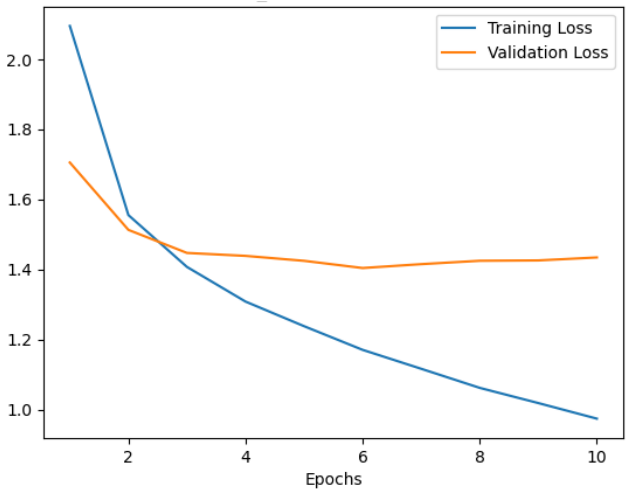


Figure 1: Training and validation loss plots for baseline LSTM

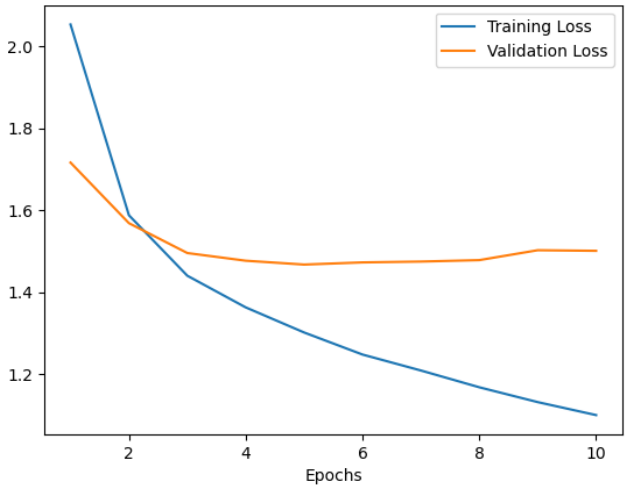


Figure 2: Training and validation loss plots for vanilla RNN

The Vanilla RNN and LSTM have very similar learning curves. This may be because vanishing or exploding gradients are not a significant factor in training, so the LSTM does not perform significantly better.

4 Cross Entropy Loss Test Set Comparison

Stochastic LSTM with default hyperparameters with a temperature of .05: 1.35
Deterministic LSTM with default hyperparameters: 1.36
Vanilla RNN with default hyperparameters: 1.37

The vanilla RNN did not perform as well as the baseline model in training loss, test loss, BLEU-1, or BLEU-4, but trained in about the same amount of time, reaching their best validation loss around three epochs.

5 Deterministic Approach Captions

LSTM:
BLEU-1: 66.67
BLEU-4: 7.95

vanilla RNN:
BLEU-1: 65.26
BLEU-4: 7.36

Our LSTM model performed slightly better than our vanilla RNN model. This may be because vanishing or exploding gradients do not make a big difference in our training, so the advantage of LSTM is not significant.

6 Baseline Model, Stochastic Approach, Varying Temperatures

Table 1: BLEU scores of varying temperatures

Temperature	BLEU-1	BLEU-4
1.5	26.89	1.22
1	48.67	2.67
.7	58.60	4.76
.2	66.08	7.89
.1	66.56	7.97
.05	66.66	8.01
.01	66.67	7.98
0 (deterministic)	66.67	7.95

A temperature of .01 appears to perform slightly better than a deterministic approach in the BLEU-4 scores, but this difference is not significant. A stochastic approach may not be significantly better than a deterministic approach because our model may not be robust enough to vary from the highest probability words.

7 Fine-tuning Embedding Size and Hidden Size

Fine-tuned model:
Hidden size: 512
Embedding size: 520
BLEU-1: 66.07
BLEU-4: 8.18

Compared with our baseline deterministic model, increasing the embedding size from 300 to 520 increased the BLEU-4 score to 8.18 but decreased the BLEU-1 score to 66.07. The plot of the model's training losses and validation losses is below. The model performs in about the same time as the baseline model.

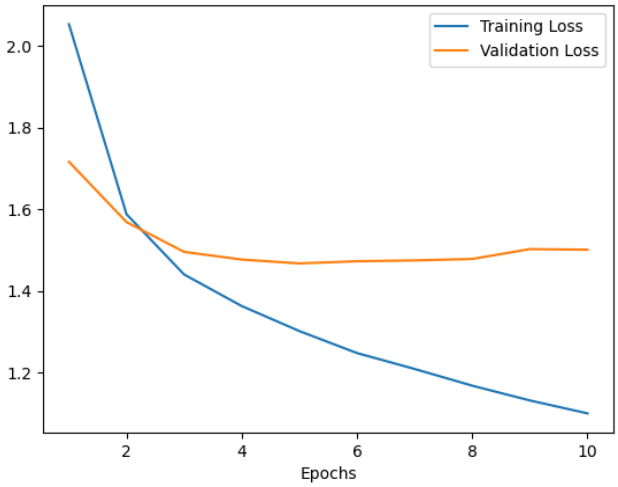


Figure 3: Training and validation loss plots for baseline LSTM with increased embed size

8 Architecture 2

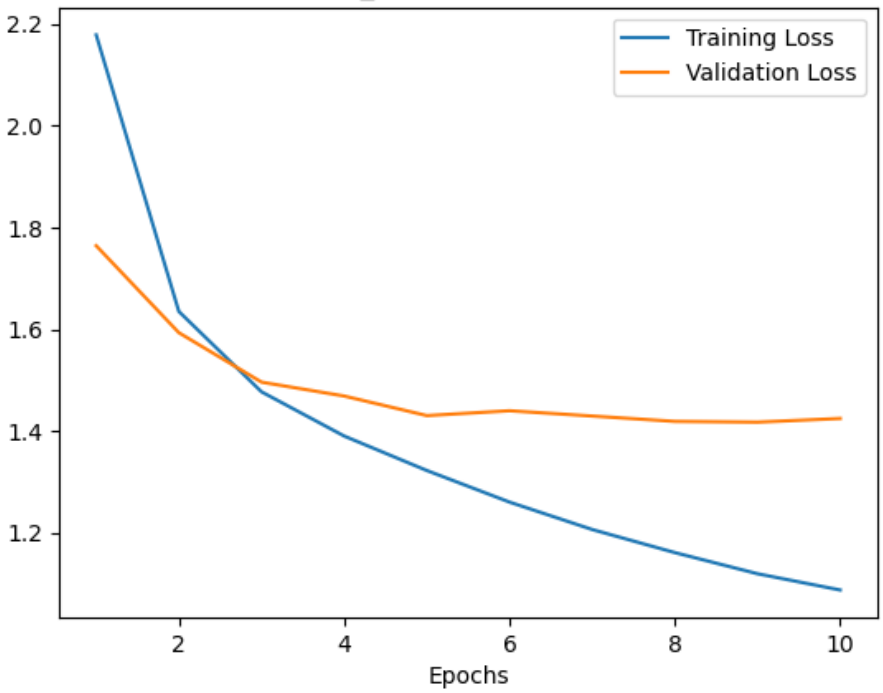


Figure 4: Training and validation loss plots for architecture 2

BLEU-1: 68.40
BLEU-4: 8.57
Architecture 2 performs better than the baseline model.

9 Captions of Best Model (Architecture 2)

Ten good captions:

Original captions:
A young man swinging a baseball bat on a baseball field.
A softball game is being played by a fence.
a person with a hat swinging a baseball bat
A young person swinging a baseball bat and striking the ball.
A boy swinging his baseball bat with a catcher and umpire behind him.
Generated caption: a baseball player is swinging a bat at a ball .

Original captions:
A tennis player has just attempted to hit the ball.

Two young men are playing tennis on a court.
A tennis player swings his tennis racquet while another person watches.
the two men are playing doubles tennis together.
A boy runs as he tries to smack the tennis ball.
Generated caption: a man holding a tennis racquet on a tennis court .

Original captions:
A clear vase that has some white flowers in it.
White flowers in a bowl of water on a table.
A bowl shaped vase with white and green flowers in it.
A vase on a table filled with flowers on a table.
A bowl filled with flowers is placed on top of a wood table.
Generated caption: a vase of flowers sitting on a table .

Original captions:
A highway sign sitting on the side of a highway.
THERE IS A HIGHWAY WITH SIGNS ON THE SIDE OF IT
A quiet highway with a street sign up ahead.

US eastbound Interstate 10 highway at Exit 352
Driving on the highway towards an exit ramp with brush on both sides.
Generated caption: a highway with a highway sign on the side of the road .

Original captions:
a red white and green fire hydrant and a fence
A fire hydrant that is on a sidewalk.
A green red and white fire hydrant on the curb.
a fire hydrant is painted red, white and green
A fire hydrant stuck in the concrete of the sidewalk.
Generated caption: a red fire hydrant on the side of a street .

Original captions:
The room has two couches in front of a tv.
The family room is clean and ready for the guests
Living room with two couches and flat screen tv over top fireplace

A coffee table stands in a living room.
A finely furnished and well-lit living room in a nice house.
Generated caption: a living room with a couch , table , and a fireplace .

Original captions:
An aeroplane soaring high in a clear sky.
A prop airplane flying in a cloudy sky.
The under side of an airplane in flight.
There is a plane flying in the sky
A small airplane is flying in the gray sky.
Generated caption: a jet flying through a cloudy blue sky .

Original caption:
a street sign with the red light glowing brightly on it
a couple of stop lights sit in front of a building
A stoplight controlling traffic in an urban intersection
An old building sits in the background behind an illuminated signal light.
A stoplight that indicates do not turn left.
Generated caption: a traffic light sitting on top of a pole .

Original captions:
A bathroom has a light above the toilet.
A white toilet sitting in a bathroom next to a bath tub.
a photograph of a bathroom with a blue wall
The bathroom is clean and crisp with a blue wall.
A small plain looking bathroom that is in someone's house.
Generated caption: a bathroom with a toilet and a sink

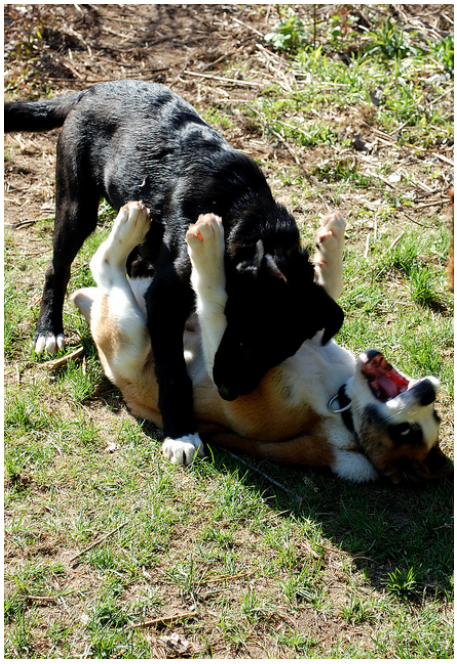
Original caption:
A delicious pizza sits on top of a white plate.
Vegetable and pepperoni pizza with a slice cut out
A pizza with pepperoni, broccoli and jalapeno peppers.
A white plate topped with a cheesy pizza.
A cooked pizza that is placed on a plate.

Generated caption: a pizza with a lot of toppings on it .

Ten bad captions:

Original captions:
a couple of people skiing down a hill
Two people in ski gear skiing down a ski slope.
A person riding ski's down a snow covered hill.
Two people on skis on slope with trees next to them.
Two people out skiing on the ski slope

Generated caption: a man in a red jacket skiing down a snow covered slope .



Original captions:

Two small dogs playing on the grass covered ground.
Two dogs having fun wrestling on a grass covered field.
Two dogs are playing with each other on green grass.
A black dog on top of brown and white dog.
Two puppies playing and rolling in the grass.

Generated caption: a dog is laying on the grass with a frisbee .

Original captions:
A woman with a suitcase holding a phone
Woman with cellphone and red suitcase standing in open area.
A women holding a red suitcase that has wheels.
A young woman stands waiting with her luggage.
a woman standing around with a red suitcase
Generated caption: a group of people standing next to each other .

Original captions:
A pizza with toppings sitting on a tray.
a small pizza sits on top of a table
A pizza placed on a large white plate.
A cheese and tomato pizza on a serving dish.
a close up of an uncooked pizza on a stove
Generated caption: a pizza with toppings sitting on a wooden table .

Original captions:
A child holding chocolate donut with both hands.
A young boy is standing against a wall eating an apple.
little boy in stripped shirt holding donut in hands
A young boy holding a doughnut with a bite out of it.
A small boy in a striped shirt eating something.
Generated caption: a little girl holding a small child 's mouth .

Original captions:
Young person practicing first aid technique with others on grassy area.
Some people gathered around a CPR mannequin to learn CPR.
The woman is learning something on a mannequin.
Some people gathered together around a dummy on a board.
A group of young people learning something outdoors.
Generated caption: a group of people sitting around a table eating .

Original captions:
A street sign depicting the corner of TUDOR CITY PLACE and E 42 ST.
A sign for East 42 street and Tudor City.
A street sign for Tudor City Place and 42 Street.
A street sign against a sky with many clouds
A street pole with street signs hanging off of it's sides.
Generated caption: a street sign that reads `` no parking "

Original captions:
A man grinding his skateboard on a rail.
A man riding a skateboard on the side of a metal rail.
A man on a skateboard grinding on a pole
A young man doing an axle grind on a piece of pipe in a park.
A boy on a skateboard rail on a skateboard
Generated caption: a person is on a skateboard in a room .

Original captions:
A surfer in a wetsuit stands with his surfboard and bicycle at the beach.
A man on a bike near a beach with a surf board on the side.
An older gentlemen with his bicycle at the beach.
A man riding his bike near the beach.
A man in skintight clothing holds a bicycle with a red surfboard.
Generated caption: a man riding a skateboard down a street .

Original captions:
The sculpture is of five people in clown costumes riding one long bike.
A statue of brightly dressed clowns riding a bicycle with multiple seats.
Five people in colorful outfits riding a long bike.
The people is dressed up as funny clowns
A statue of clowns riding a tandem bike,
Generated caption: a man riding a bike past a group of people .

12 Authors' Contributions

Table 5: Authors' contribution

Noah Inada	worked on the encoder and decoder, deterministic and stochastic functions, report, fine-tuning
Kaiwen Tan	worked on architecture 2, worked on building a model for optimizer, generated captions and image for report
Sheung Ho	worked on bleu score, worked hard on fine-tuning parameters, worked hard on vanilla RNN, worked hard on report,

13 References

pytorch.org

piazza.com

https://towardsdatascience.com