Cuckoo For COCO Captions: Image Description with CNNs and LSTMs

Rye Gleason

Benson Duong

Takuro Kitazawa

Jeremy Nurding

Abstract

In this project, we trained an image captioner with CNN (modified alexnet or pre-trained resnet) and LSTM, and did hyper-parameter fine tuning, changing hidden size. We found that pre-trained yields better results. In addition, when generating with temperature (which divides the softmax), we saw that higher values of temperatures worsen the resulting captions.

1 Introduction

The goal of the project is to generate text captions given images. The approach used is getting a traditional convolutional neural network to embed 2D images into fully connected vectors which will hold latent representation of the features and elements in the image; this is then fed into a recurrent neural network using LSTM's, with (already provided captions that are also embedded and put in every time step following the first image), that then predicts the sequence of words. By doing so, images alone can be fed first into the model and output predicted text captions. 2 architectures were used for the preliminary CNN, which were a modified alexnet, and a pretrained resnet. PyTorch was used. Further experimentation was done with hyper-parameter fine-tuning on the hidden size hyperparameter.

2 Background and Related Work

Resources included the pretrained Resnet (`https://pytorch.org/vision/stable/models.html`), the COCO 2015 Image Captioning Task (`https://cocodataset.org/`) which has a dataset with images and corresponding captions. Pytorch documentation pages include those for implementing pretrained weights (`https://d2l.ai/chapter_computer-vision/fine-tuning.html`) , and LSTM layer documentation and teacher forcing:

`https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html`,
`https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html`

Two previous papers using COCO are "Improved Image Captioning via Policy Gradient Optimization of SPIDEr" and "Rich Image Captioning in the Wild," which use COCO as a benchmark for testing SPIDEr, a replacement for BLEU; and several new advanced imaged captioning models, respectively.

Table 1: CNN Layers

Layer	Input Channels	Output Channels	Stride Size	Kernel Size	Activation	Padding Size
conv1	3	64	4	11	ReLU	0
maxpool1	64	64	2	3	Identity	0
conv2	64	128	1	5	ReLU	2
maxpool2	128	128	2	3	Identity	0
conv3	128	256	1	3	ReLU	1
conv4	256	256	1	3	ReLU	1
conv5	256	128	1	3	ReLU	1
maxpool3	128	128	2	3	Identity	0
adaptive avgpool	128	128	1	1	Identity	0
fully connected 1	128	1024	N/A	N/A	ReLU	N/A
fully connected 2	1024	1024	N/A	N/A	ReLU	N/A
fully connected 3	1024	300	N/A	N/A	Identity	N/A

3 Models

The model used for this project was a Convolutional Neural Network (CNN) encoder combined with a Long Short-Term Memory (LSTM) decoder. The CNN used is a variant of AlexNet, and its architecture is shown in table 1. The LSTM used has 2 layers with a hidden size of 512 units, an embedding size of 300 units. It uses biases, and has no dropout. Finally, in some experiments we replaced our custom CNN with a pretrained ResNet. The ResNet architecture is best known for introducing "skip" gates, which allow for each layer of the network to be calculating only residuals, not the entire new activation value. This helps deal with the vanishing/exploding gradient problem.

For our experiments, we decided to vary the number of hidden units in the LSTM. We were curious if 512 units was overkill, and if the LSTM could perform as well with fewer units, and therefore fewer weights, so we also tried training a version of the network with 256 hidden LSTM units. This gives as a total of 4 models: our custom CNN with a 512-unit LSTM, called "custom 512," ResNet with a 512-unit LSTM, called "ResNet 512," our custom CNN with 256 hidden units, called "custom 256," and ResNet with 256 hidden units, called "ResNet 256."

4 Results

BLEU scores and test losses for all models are shown in table 2. A possible reason why Custom CNN didn't work as well as ResNet was because the custom was trained from scratch whereas the resnet was already provided with pre-trained weights, possibly optimized on GPUs and training time far better than any of our teammates could muster. There are also specific qualities of resnet that give it advantageous results over other architectures, namely its use of residual skip connections during backpropagation. Both of these models outperformed their counterparts since having a higher hidden size means more weights and in turn more degrees of freedom to learn the model.

Table 2: BLEU Scores

Name	BLEU-1	BLEU-4	Test Loss
Custom 512	45.53%	1.47%	1.38
Custom 256	45.03%	1.43%	1.40
Resnet 512	62.88%	5.01%	1.34
Resnet 256	61.9%	4.29%	1.48

The best custom model, custom 512, has a test loss of 1.38, and the test loss of the best ResNet model, ResNet 512, is 1.34. The training and validation loss curves for those models are shown in figures 1 and 2, respectively. Both models have a rather wide difference by the 10th epoch. Another similarity is that the training loss dips below the validation loss by the 2nd epoch, or roughly around it. One difference is that the pre-trained model's validation loss is less smooth than the Custom CNN's validation loss. Another difference of course is that the custom 512 model has a higher validation loss than the ResNet 512 model.

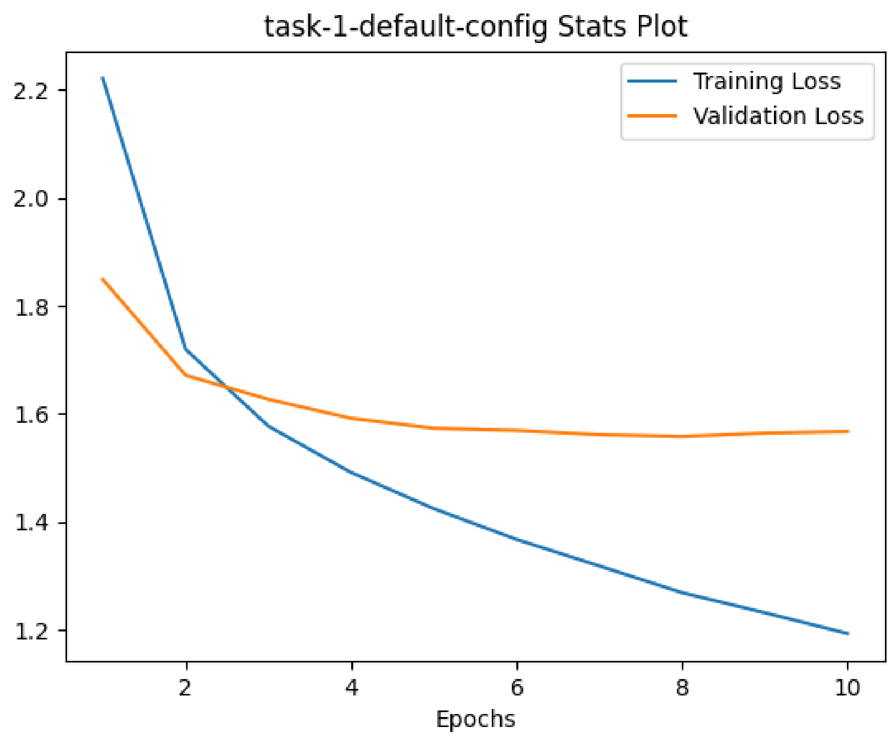


Figure 1: Training and validation loss curves for the "Custom 512" model, the better-performing custom model.

5 Captions

See figures.

6 Discussion

Our results show that ResNet preformed much better than our Custom CNN. One reason ResNet outperforms our CNN is because it overcomes the vanishing gradient problem by creating multiple layers, skips layers, and reuses previous layers. By overcoming the vanishing gradient problem, ResNet is able to make a model with thousands of convolutional layers. Meanwhile, our Custom CNN only uses 5 convolutional layers. For our Custom CNN, increasing the number of hidden layers improves the performance of our model. However, once the number of hidden layers becomes greater than the optimal number of hidden layers, the time complexity of our model increases faster than the accuracy gained by the model.

7 Team Contributions

7.1 Rye

Implemented ResNet CNN. Did debugging.

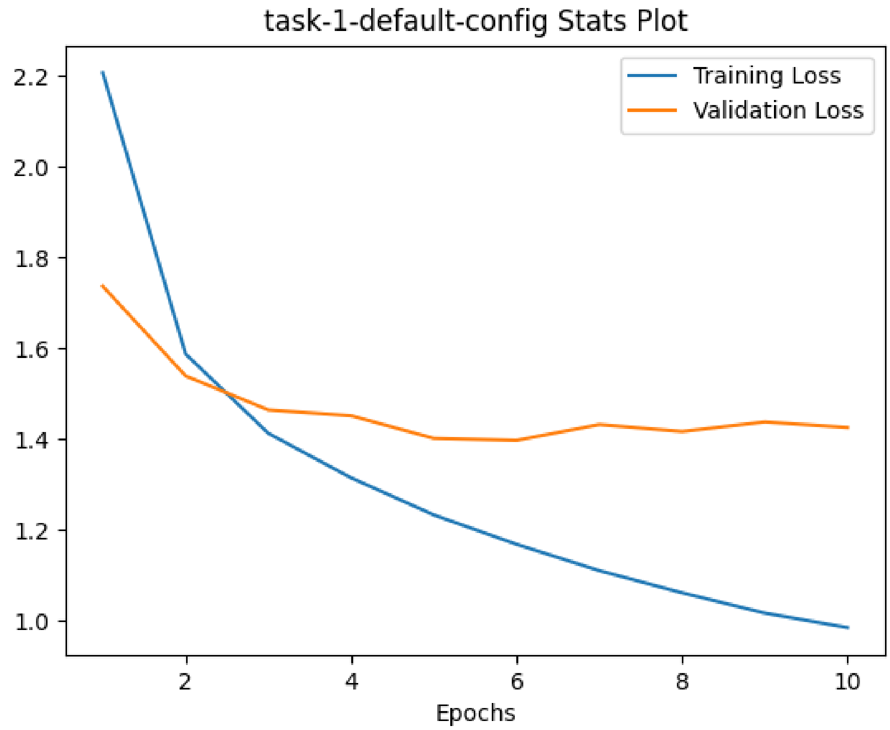


Figure 2: Training and validation loss curves for the ”ResNet 512” model, the better-performing ResNet model.

7.2 Benson

Worked on Custom CNN. Worked on portions of LSTM (the non-teacher forcing part). Did debugging.

7.3 Takuro

Worked on LSTM except the non-teacher forcing part. Did debugging.

7.4 Jeremy

Worked on experiment.py. Ran Custom CNN model with lower hidden size. Wrote discussion section in report.

Figure 3: ”Good” captions generated by Custom 512 at various temperatures:
0.4: a bathroom with a sink and a toilet
5: if magazine roll chipped and ovens bag bakers license city can scrubland baseball attire tad eaten computer vending virgin permitting
0.001: a bathroom with a sink, toilet and a shower.
Deterministic: a bathroom with a sink , toilet and a shower.

Figure 4: ”Good” captions generated by Custom 512 at various temperatures:
0.4: a man in at a table with a plate of food
5: possible sideways icebox activities during shake checkerboard clocks pittsburgh overlooking ti mariner filling manuals pipe fan contrasting checks cancer
0.001: a man is holding a banana in his hand.
Deterministic: a man is holding a banana in his hand.

Figure 5: ”Good” captions generated by Custom 512 at various temperatures:
0.4: a woman man riding a surfboard on a wave in the ocean.
5: hotels connect brim jumping reach northern tried nathans demonstration stack hood router mane log linger fries somewhat dispensers swing rested
0.001: a man on a surfboard riding a wave.
Deterministic: a man is riding a horse in a field

Figure 6: ”Bad” captions generated by Custom 512 at various temperatures:
0.4: a woman holding a baseball racquet at a tennis ball.
5: mark beaming and controlled landscape sideways booth 1960s paneled whizzes eagles threw bumper-to-bumper corner crowded results phillie 1897 strings horizontal
0.001: a man is riding a motorcycle on the street.
Deterministic: a man is riding a motorcycle on the street.

Figure 7: ”Bad” captions generated by Custom 512 at various temperatures:
0.4: a man is sitting a a frisbee wii controller a living room .
5: kitchen artichokes establishments boogey teddybears horizontal seatbelt country engines conversations microwave touches grafitti doors antelope couple vigorously ribbon candid yacht
0.001: a man is holding a banana in his hand.
Deterministic: a man is holding a cell phone in his hand.

Figure 8: ”Bad” captions generated by Custom 512 at various temperatures:
0.4: a woman in a black is tie standing standing. holding man is talking a phone phone.
5: weds , nutella wearing pick ramekin glare itching unique chickpeas vendor colorless everywhere satchel observed crinkle seperating beginning severed sizes
0.001: a man in a suit and tie standing in front of a tv.
Deterministic: a man in a suit and tie standing in front of a tv.

3 Images with Good captions at temp = 0.4

ID = 131453

Temp = 0(Deterministic)
a herd of elephants walking across a dirt field .

Temp = 0.001
a herd of elephants walking across a dirt field .

Temp = 0.4
a herd of elephants standing across adrift field

Temp = 5
bikes adventure with grassy armchairs sunny wetsuit spray clorized listen ta charlottesville celebrations bull quaint tug tank up momentos polymer

ID = 14297

Temp = 0(Deterministic)
a boat is parked on the shore of a river .

Temp = 0.001
a boat is parked on the shore of a river .

Temp = 0.4
a boat on a boat on the water and a boat on the water .

Temp = 5
suede vegetation enjoying partial reindeer together hosing drift dwarfs capped ex back toward 28 thin sails i updated leafs available

ID = 167894

Temp = 0(Deterministic)
a street sign that says `` `` < unk > `` `` on a pole .

Temp = 0.001
a street sign that says `` `` < unk > `` `` on a pole .

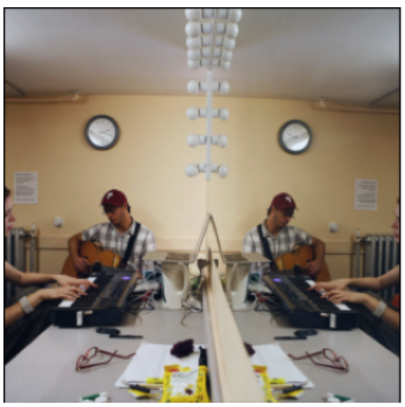
Temp = 0.4
a street sign that says on a pole .

Temp = 5
king frisby ; tki shows erected aged flowered stone monarch tried motorcycle arrows flush implements h completes looming entering creates

Figure 9: Images with corresponding ”good” captions generated by ResNet 512 at various temperatures.

3 Images with Bad captions at temp = 0.4

ID = 10040
Temp = 0(Deterministic)
a man and a woman are standing in a parking lot .
Temp = 0.001
a man and a woman are standing in a parking lot .
Temp = 0.4
a man of people standing around front of a building .
Temp = 5
suffolk elder fire brother donkey fatigues rink / draining leaping male manipulating b barriers ferris saute curling five ion biled



ID = 104393
Temp = 0(Deterministic)
a man is standing in a room with a toothbrush .
Temp = 0.001
a man is standing in a room with a toothbrush .
Temp = 0.4
a woman is wearing ready to be picked ball cake .
Temp = 5
lounge skiers catering monte greets kitties chairs coach looking lonely leaks mini-fridge i bathing park tested broad songbird idle clamp

ID = 205108
Temp = 0(Deterministic)
a man is eating a slice of pizza from a pan .
Temp = 0.001
a man is eating a slice of pizza from a pan .
Temp = 0.4
a man is holding a pizza of pizza on
Temp = 5
egg secured headlights classy covered customer bib lining menu watercraft kg bureau hitched cardboard freeze work objects specialized siscs fixins

Figure 10: Images with corresponding ”bad” captions generated by ResNet 512 at various temperatures.