Mix and Match: Image Caption Generators

Kwan Kiu Choy
University Of Toronto
choykwan@mail.utoronto.ca

Parinita Edke
University of Toronto
parinita.edke@mail.utoronto.ca

Natalie Ashgriz
University of Toronto
n.ashgriz@mail.utoronto.ca

Abstract

The generation of image captions has attracted the attention of researchers in the field of AI due to its many applications in a variety of fields. In this paper, we explore the impacts of substituting the vision deep CNN component of the caption generation model presented by Vinyals et al with other pre-trained models. We experiment with AlexNet, VGG-19, and Inception V3 for the vision component of the model and the LSTM model from the Vinyals et al paper for the language generating component of the model and see if there is a difference in performance in caption generation. We apply these modified models to the Flickr8k dataset and compare our models' performances using BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores. We find that AlexNet outperforms VGG-19 and Inception V3 when trained with a lower learning rate.

1 Introduction

The generation of image captions has gradually attracted the attention of researchers in the field of AI due to its applications in a variety of fields. Generating image captions involves the automatic generation of natural language descriptions based on the content observed in an image. It combines the knowledge of both computer vision and NLP, playing an important part in scene understanding and analysis. Image caption generation is very important to making digital media accessible to those who use screen readers. In this paper, we will explore the impacts of substituting Inception V3, AlexNet, and VGG-19 into the GoogleLeNet CNN component of the caption generation model presented by Vinyals et al.

2 Related Works

2.1 Caption Generating Neural Network Models

The computer vision community has been interested in generating natural language descriptions for images and videos for a long time [4]. In this paper, we focused on the Google NIC model proposed by Vinyals et al, which is a combination of a CNN (GoogLeNet) with weights pre-trained in ImageNet for detecting objects, and an LSTM for generating the captions [1]. The model achieved better BLEU-1 scores than the previous state-of-the-art model for four of the datasets used, with improvements of 136%, 17% 8%, 47% respectively. Moreover, the model achieved better performance than humans on the MSCOCO dataset by 27%. While the Google team focused on a more sophisticated sentence generation model, Li and Karpathy et al proposed using a simpler multimodal RNN to complete the same task and a RCNN for object detection [2]. Both models perform similarly. Inspired by their work, we compare the performance of models consisting of different combinations of CNN/RCNN that are pretrained in ImageNet with an LSTM, and then pick the best one to form our model.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

3 Methods

3.1 Flick8k Dataset

The Flick8k dataset is a dataset that is designed for image captioning purposes. It contains 8000 images and for each image, 5 captions are provided. This is widely used for image captioning tasks, and it is often referred to as the beginner dataset due to its relatively small size compared to other common datasets such as MSCOCO and Flickr30k. The train-validation-test splits were taken from Hodosh et al [10]. We use the words in the captions to develop our vocabulary.

3.2 Preprocessing - Image Transformation and Text Tokenization

We first resize the images to predefined sizes. Then, we convert the images to tensors and normalize them with mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]. These mean and standard deviation values are obtained from ImageNet. For captions, we clean the text by removing extra spaces and punctuation, and convert all the sentences to lowercase. Afterward, we split the captions into a list of words and tokenize them using spacy in order to pass them into the embedding layer.

3.3 Model

Our model adopts an encoder-decoder architecture, where the images are first passed into an image encoder so that features embeddings will be returned. The feature embeddings are passed to the decoder model and generate the natural language captions for the images.

3.3.1 Image Encoder and Transfer Learning

For our image encoder, we decided to use pre-trained models from ImageNet. We wanted to compare the performance between AlexNet [7] , VGG-19 [8] and Inception V3 [9]. Since these models were originally designed for classification purposes, there were slight modifications made to the last classification layer. The modifications are documented below. In the training process, we also fine-tune the last layers of the pre-trained model to get higher accuracy.

For Inception V3, we disable the auxiliary net. Apart from that, we change the fully connected layer into a linear layer so that it will output a feature vector of embedding size at the end of the forward pass. The changes that were made to AlexNet and VGG-19 are similar. The last layer of the original AlexNet and VGG-19 net is a linear layer that outputs a feature vector of size 1000. We change the outputting size from 1000 to embedding size in the last layer for our purposes.

3.3.2 Sentence Decoder

We pass in the image features and embedded captions (as described in 3.2) into the LSTM network, and obtain an output of size 256 (the hidden size). The associated equations for the LSTM network are listed as below.

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{1}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{2}$$

$$g_t = tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{3}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

The output of the decoder is passed into a linear layer that outputs a vector with indices that corresponds to different words in our vocabulary.

4 Results

In order to find the optimal model, we run experiments on the three different models using various hyperparameter settings. We try tuning the batch size [128, 256], the learning rate [3e-2, 3e-4], and the number of layers [1, 2] in the LSTM network. For each of the experiments, we run the model for 50 epochs, with hidden size and embedding size being 256 and 512, respectively.

4.1 Selected Test Images

Below are the images we use to test the trained models. Each image has 5 captions labelled by humans, one of which is provided in the image caption. The rest can be found in Appendix A.

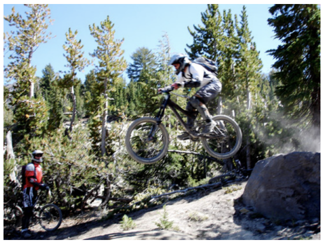

Figure 1: A helmeted man jumping off a rock on a mountain bike.

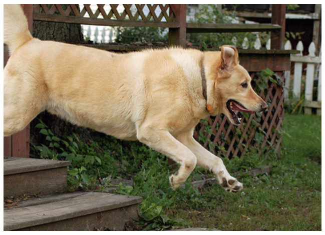

Figure 2: A blond dog runs down a flight of stairs to the backyard

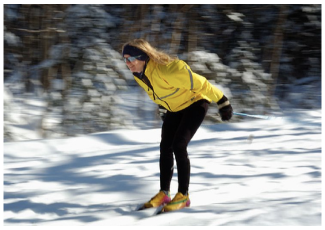

Figure 3: A woman in a yellow and black outfit is skiing

4.2 Hyperparameter settings 1

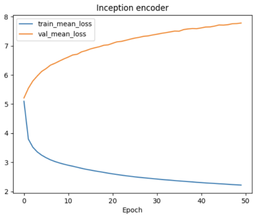

Figure 4: The training and validation loss for the Inception V3 captioning model

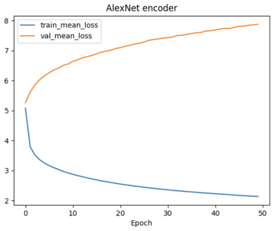

Figure 5: The training and validation loss for the AlexNet captioning model

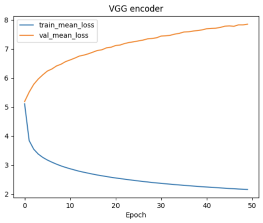

Figure 6: The training and validation loss for the VGG-19 captioning model

The validation loss and training loss start at about the same value, however quickly diverge. Although this is often a sign of overfitting, we notice that our predictions (see Appendix B) improved qualitatively as the number of epochs increased. This would not be the case if the model was overfitting, so we are not sure what is causing the validation loss to increase.

The best captions (selected qualitatively) that are generated from this series of experiments are:
Figure 1: a man in a red shirt is riding a bike on a dirt path (AlexNet)
Figure 2: a dog is running through the grass (VGG-19)
Figure 3: a man in a red jacket is skiing down a snowy hill (VGG-19)

4.3 Hyperparameter settings 2

This set performed poorly. The resulting captions (see Appendix C) consist of the same sentence, regardless of the input image. We believe that the learning rate is too high and as a result, the model gets stuck in an ill-conditioned curve, causing it to oscillate perpetually instead of continuing to learn. This makes sense since the training loss drops initially but plateaus very quickly.

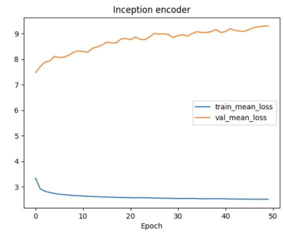

Figure 7: The training and validation loss for the Inception V3 captioning model

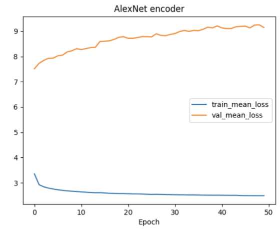

Figure 8: The training and validation loss for the AlexNet captioning model

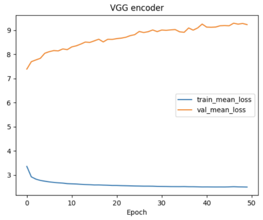

Figure 9: The training and validation loss for the VGG-19 captioning model

The best captions (selected qualitatively) that are generated from this series of experiments are:
Figure 1: a man in a red shirt and jeans is standing on a rock (AlexNet)
Figure 2: a man in a blue jacket is sitting on a bench (Inception V3)
Figure 3: a man in a red shirt is playing with a racquet (VGG-19)

4.4 BLEU Scores

Model	Hyperparameter setting	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Inception	1	0.4066	0.1575	0.2896	0.3863
Inception	2	0.5455	0.2335	0.3829	0.4833
AlexNet	1	0.5451	**0.6463**	**0.6917**	**0.7438**
AlexNet	2	0.5385	0.2996	0.4513	0.5473
VGG-19	1	0.5495	0.3728	0.3129	0.3989
VGG-19	2	**0.5714**	0.2965	0.1974	0.2926

Table 1: Average BLEU scores for 5 captions generated for image in Figure 1

We compute the BLEU scores by generating 5 candidate captions for the image in Figure 1 using each of the six models, and using the 5 captions from the Flickr8k dataset as the references. Note that the average score is not robust to fluctuations in individual BLEU scores.

5 Discussions and Future Directions

After the series of experiments, we find that AlexNet performs the best, followed by VGG-19, then Inception V3. The BLEU scores we report are especially high for AlexNet. The quality of the predicted captions in hyperparameter setting 1 are also reasonable, even though AlexNet does not identify colours well.

A problem that we find is that the BLEU score may not be the best metric for evaluating the performance of our models. We conduct a small experiment on this and find interesting results. When the valid caption is “a dog running on a beach”, and we try “a dog is running” and “a cat is running”, the BLEU scores are very similar to each other. Callison-Burch et al stated that the problem with the BLEU score is its acceptance of a large variety of hypotheses, which does not consider any grammatical or semantic errors.[6] This is an issue, since in caption generation, there may be many valid captions that do not use the exact same words as the references.

For future work, we would like to see how our models perform on larger datasets, such as MSCOCO and the Google Conceptual Captions dataset. Training the model with more images will likely help generate more reasonable captions. We would also like to test whether adding attention [5] to image encoder would give us better results.

6 Summary

To summarize, our project explores how using different vision models for image captioning will vary the performance of the caption generator. We find that AlexNet and VGG-19 perform similarly well, while the Inception V3 net (with auxiliary net disabled) performs worse. The Inception V3 net's poor performance may be caused by the small number of epochs that we ran it with. We believe that with more computing resources, there could be a significant increase in its BLEU score. While working on the project, the question of whether BLEU score is a suitable metric for evaluating the generated captions was also raised, and we believe that this is a topic where we can delve deeper in the future. For enhancements, we would like to apply our model to larger datasets, such as MSCOCO and Flickr30k, as well as add attention to the image encoder.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. http://arxiv.org/abs/1411.4555

[2] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Trans. Pattern Anal. Mach. Intell. 39, 4 (April 2017), 664–676. DOI:https://doi.org/10.1109/TPAMI.2016.2598339

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. DOI:https://doi.org/10.1145/3065386

[4] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. https://aclanthology.org/P18-1238.pdf

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057, 2015.

[6] Chris Callison-Burch Miles Osborne Philipp Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. link: https://aclanthology.org/E06-1032.pdf

[7] Alex Krizhevsky. 2014. One weird trick for parallelizing convolutional neural networks. CoRR. abs/1404.5997 DOI: http://arxiv.org/abs/1404.5997

[8] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. DOI: https://doi.org/10.48550/arXiv.1409.1556

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. 2015. CoRR. abs/1512.00567. DOI:http://arxiv.org/abs/1512.00567

[10] M. Hodosh, P. Young and J. Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artifical Intellegence Research, Volume 47, pages 853-899

7 Contribution

7.1 Kwan Kiu Choy

Worked on some code, Related Works, Methods (Flick8kDataset, Preprocessing, Model) Discussions and Future Directions

7.2 Parinita Edke

Worked on code, ran the models due to having a local GPU, generated visualizations, worked on introduction and analysis.

7.3 Natalie Ashgriz

Worked on code, ran the models using GPU, analysis of results, calculation of BLEU scores

8 APPENDIX

8.1 Appendix A

Caption	Figure 1
1	A helmeted man jumping off a rock on a mountain bike .
2	A man jumping on his bmx with another bmxer watching .
3	A mountain biker is jumping his bike over a rock as another cyclist stands on the trail watching .
4	A person taking a jump off a rock on a dirt bike .
5	The bike rider jumps off a rock .

Table 2: Label captions for image in Figure 1

Caption	Figure 2
1	A blond dog runs down a flight of stairs to the backyard .
2	A dog jumps off the stairs .
3	A tan dog runs down a wooden staircase to the green grass .
4	A yellow dog is jumping across a grassy yard in front of a wooden deck .
5	A yellow dog jumping off of a porch .

Table 3: Label captions for image in Figure 2

Caption	Figure 3
1	A woman in a yellow and black outfit is skiing .
2	A woman skier in yellow and black races down the slope in front of the snow-covered trees .
3	A woman skiing down a slope .
4	A woman snow skiing in a yellow jacket and black pants .
5	The woman in the yellow jacket is riding on snow skates .

Table 4: Label captions for image in Figure 3

8.2 Appendix B

Caption	Figure 1
1	a man in a blue shirt and a white shirt is standing on a sidewalk with a dog
2	a man in a white shirt and a white shirt and a black shirt is standing on a sidewalk
3	a man in a black shirt and a white shirt is standing on a sidewalk with a <UNK>
4	a man is standing on a <UNK> with a <UNK> in the background
5	a man in a blue shirt and a white shirt is standing on a sidewalk with a <UNK>

Table 5: Captions predicted by InceptionV3 using Hyperparameter 1 for image in Figure 1

Caption	Figure 1
1	a man in a red jacket is riding a bike in the woods
2	a man in a red jacket is riding a bike in the woods
3	a man in a red shirt is riding a bike on a dirt path
4	a person in a red jacket is riding a bike in the woods
5	a man in a red jacket is riding a bike in the woods

Table 6: Captions predicted by AlexNet using Hyperparameter 1 for image in Figure 1

Caption	Figure 1
1	a man in a red shirt is riding a bike on a dirt path
2	a man in a red shirt is riding a bike on a dirt path
3	a man in a red shirt is riding a bike on a dirt path
4	a man in a red shirt is riding a bike on a dirt track
5	a man in a red helmet is riding a bike down a hill

Table 7: Captions predicted by VGG-19 using Hyperparameter 1 for image in Figure 1

Caption	Figure 2
1	a woman in a black and white dress is walking down a sidewalk
2	a man in a black shirt and a white shirt is standing on a sidewalk with a large brown dog
3	a man in a black shirt and a white shirt is standing on a sidewalk with a <UNK>
4	a man is standing on a bench with a white dog
5	a man and a woman are sitting on a bench

Table 8: Captions predicted by InceptionV3 using Hyperparameter 1 for image in Figure 2

Caption	Figure 2
1	a dog is walking on the grass
2	a brown and white dog is running through a grassy area
3	a brown dog is standing on a leash
4	aa brown dog is standing on a <UNK> carpet
5	a brown dog is standing on a <UNK> carpet

Table 9: Captions predicted by AlexNet using Hyperparameter 1 for image in Figure 2

Caption	Figure 2
1	a dog is standing in the snow
2	a brown dog is running through a grassy area
3	a dog is standing on a brown horse
4	a dog is running through the grass
5	a dog is standing on a brown horse

Table 10: Captions predicted by VGG-19 using Hyperparameter 1 for image in Figure 2

Caption	Figure 3
1	a man in a black shirt and a white shirt is standing on a sidewalk with a large brown dog
2	a man in a black shirt and a white shirt is standing on a sidewalk with a <UNK>
3	a man in a black shirt and a white shirt is standing on a sidewalk with a large brown dog
4	a little girl in a pink dress is playing with a ball in the grass
5	a man in a black shirt and a white shirt is standing on a sidewalk with a <UNK>

Table 11: Captions predicted by InceptionV3 using Hyperparameter 1 for image in Figure 3

Caption	Figure 3
1	a young girl in a pink jacket is climbing a rock
2	a man in a red shirt and a blue shirt is playing with a ball in the sand
3	a man in a yellow shirt is standing in the snow
4	a man in a yellow shirt is standing on a beach
5	a young boy in a blue shirt is running through the snow

Table 12: Captions predicted by AlexNet using Hyperparameter 1 for image in Figure 3

Caption	Figure 3
1	a man in a red jacket is snowboarding down a snowy hill
2	a skier is jumping over a snowy hill
3	a man in a red jacket is standing on a snowy hill
4	a man in a red jacket is skiing down a snowy hill
5	a man in a red jacket is standing on a snowy hill

Table 13: Captions predicted by VGG-19 using Hyperparameter 1 for image in Figure 3

8.3 Appendix C

Caption	Figure 1
1	a man in a blue jacket is sitting on a bench
2	a man in a blue jacket is sitting on a bench
3	a man in a blue jacket is sitting on a bench
4	a man in a blue jacket is sitting on a bench
5	a man in a blue jacket is sitting on a bench

Table 14: Captions predicted by InceptionV3 using Hyperparameter 2 for image in Figure 1

Caption	Figure 1
1	a man in a red shirt and jeans is standing on a rock
2	a man in a red shirt and jeans is standing on a rock
3	a man in a red shirt and jeans is standing on a rock
4	a man in a red shirt and jeans is standing on a rock
5	a man in a red shirt and jeans is standing on a rock

Table 15: Captions predicted by AlexNet using Hyperparameter 2 for image in Figure 1

Caption	Figure 1
1	a man in a red shirt is playing with a racquet
2	a man in a red shirt is playing with a racquet
3	a man in a red shirt is playing with a racquet
4	a man in a red shirt is playing with a racquet
5	a man in a red shirt is playing with a racquet

Table 16: Captions predicted by VGG-19 using Hyperparameter 2 for image in Figure 1

Caption	Figure 2
1	a man in a blue jacket is sitting on a bench
2	a man in a blue jacket is sitting on a bench
3	a man in a blue jacket is sitting on a bench
4	a man in a blue jacket is sitting on a bench
5	a man in a blue jacket is sitting on a bench

Table 17: Captions predicted by InceptionV3 using Hyperparameter 2 for image in Figure 2

Caption	Figure 2
1	a man in a red shirt and jeans is standing on a rock
2	a man in a red shirt and jeans is standing on a rock
3	a man in a red shirt and jeans is standing on a rock
4	a man in a red shirt and jeans is standing on a rock
5	a man in a red shirt and jeans is standing on a rock

Table 18: Captions predicted by AlexNet using Hyperparameter 2 for image in Figure 2

Caption	Figure 2
1	a man in a red shirt is playing with a racquet
2	a man in a red shirt is playing with a racquet
3	a man in a red shirt is playing with a racquet
4	a man in a red shirt is playing with a racquet
5	a man in a red shirt is playing with a racquet

Table 19: Captions predicted by VGG-19 using Hyperparameter 2 for image in Figure 2

Caption	Figure 3
1	a man in a blue jacket is sitting on a bench
2	a man in a blue jacket is sitting on a bench
3	a man in a blue jacket is sitting on a bench
4	a man in a blue jacket is sitting on a bench
5	a man in a blue jacket is sitting on a bench

Table 20: Captions predicted by InceptionV3 using Hyperparameter 2 for image in Figure 3

Caption	Figure 3
1	a man in a red shirt and jeans is standing on a rock
2	a man in a red shirt and jeans is standing on a rock
3	a man in a red shirt and jeans is standing on a rock
4	a man in a red shirt and jeans is standing on a rock
5	a man in a red shirt and jeans is standing on a rock

Table 21: Captions predicted by AlexNet using Hyperparameter 2 for image in Figure 3

Caption	Figure 3
1	a man in a red shirt is playing with a racquet
2	a man in a red shirt is playing with a racquet
3	a man in a red shirt is playing with a racquet
4	a man in a red shirt is playing with a racquet
5	a man in a red shirt is playing with a racquet

Table 22: Captions predicted by VGG-19 using Hyperparameter 2 for image in Figure 3