Image Captioning on Flickr8K using ResNet101 and Bahdanau Attention

Abhay Zala Garv Srivastava Meghana Sankaran David Karash

***Abstract*—Image captioning is a method receiving increasing attention in the AI space and is a particularly relevant area of research that continues to advance. In this paper, we explore an implementation of an image captioning model and how various modifications effect its performance on the task. We test with a simple encoder-decoder architecture using the pretrained ResNet101 model and an LSTM. We evaluate our models under multiple industry standard metrics, including Bleu, METEOR, ROUGE_L, CIDEr, and SPICE. Limited success was achieved by the models, which leaves space to question why these models underperform against other similar models when using ResNet101 that should outperform similar models using VGG. Hyperparameter tuning was also not utilized which may contribute to the poor performance of our models.**

***Index Terms*—machine learning, image captioning, captioning, computer vision, natural language processing**

I. Introduction

Image captioning is a method receiving increasing attention in the AI space and is a particularly relevant area of research that continues to advance. In this paper, we display the implementation of image captioning to infer the semantics of an input image and produces a sufficiently detailed description for the objects, scenes, and context within the image. To do so, we test several iterations of an encoder-decoder architecture with ResNet101 and LSTM, which allows for a more rigorous investigation into the synergetic performance of the two models. This approach illuminated the nuances of some of the predominant techniques in image captioning. Our code can be found at: https://github.com/aszala-UNC/COMP562-final-project.

A. Motivation

Image Captioning is a cutting-edge field within Deep Learning that aims to capture the relationship between objects present in images and generate descriptions to describe these relationships. The process uses Natural language Processing and Computer Vision to generate these captions. Through the use of very powerful neural networks and other machine learning techniques, data scientists are able to encode images within high dimension vector spaces to extract certain features, then use NLP processes to decode these features and understand them as words that the model is trained with. Image Captioning has gained a lot of traction in our world today, and various companies are implementing this technology for accessibility and service products. As an example, NVIDIA is using image captioning technologies to create an application to help people with reduced or no eyesight. Image captioning also has tremendous upside within the medical field, where nurses and doctors can leverage sophisticated image captioning technologies to rapidly detect the presence of certain diseases or irregularities within patients.

Fig. 1. The image on the left is captioned: "A child in a pink dress is climbing up a set of stairs in an entry way". The image on the right is captioned: "A black dog and a tri-colored dog playing with each other on the road".

B. Related Work

Generating natural language descriptions from visual data has long been studied in computer vision. This has led to complex systems composed of visual primitive recognizers combined with a structured formal language, such as And-Or Graphs and logic systems, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively brittle and have been demonstrated only on limited domains, e.g. traffic scenes or sports. The approach taken in this paper has had significant impact on major technological applications. A prominent example is in 2014, when researchers from Google developed a state-of-the-art neural image caption generator on the MSCOCO dataset. The model utilized Convolutional Neural Networks and Long-Short Term Memory (LSTM) to produce extremely accurate results, some of which can be seen below.

They used CNN to develop a dense feature vector, which served as their embedding process for the images. It helps to think of this feature vector as a nearly direct translation of a language that the model cannot understand. The density of these feature vectors depends on the complexity of the image itself. This feature vector serves as the initial state for the LSTM.

Fig. 2. A subset of image and their generated captions, grouped by the people of Google.

To get a basic understanding of the purpose of LSTM, it is helpful to think of our own ways of understanding: Humans do not have thoughts that are developed on the spot. When reading something, the meaning of each word is understood based on the meaning of the previous ones. There is the persistence of memory, a long-term one that helps you understand the overall context of what you are reading, and a short-term one that helps each word make sense with the previous ones. LSTM networks are a special form of neural networks that address the issue of long-term and short-term information persistence.

Furthermore, the usage of ResNet and similar methods has a clear incision in the field of image captioning, demonstrated by several academic works implementing variants of the methods in this study. For example, an earlier implementation of ResNet is presented in Bhatia et al. 2019, wherein a standard CNN and RNN encoder is used to label image content. Atliha et al. 2022 compares the performance of the standard deep CNN architecture in VGG and ResNet as two encoders, and finds that ResNet outperforms VGG with fewer training epochs; this is due to its ability to solve the "vanishing gradients" issue found in many architectures with a higher number of layers. Thus, a further exploration of ResNet encoding is necessary, especially when implemented in conjunction with the addition of the dropout function and Bahdanau attention. Both additions have also informed academia in image captioning; a prominent example includes Al-Malla et al. 2022, which shows that Bahnadau attention system makes CNN model more smooth and differentiable.

Several other approaches have historically been taken for image captioning, including local binary patterns (Ojala et al. 2000), scale-invariant features transform (David et al. 2004), histogram of oriented gradients (Dalal et al. 2005) and a variety of other techniques to extract features from input data. Classifiers such as SVM can also be used to categorize objects within an image. However, such methods are largely handcrafted, making them less feasible with a large and diverse set of data. Other image encoders, such as AlexNet, VGGNet, GoogLeNet, and Inception-V3 (Hossain et al. 2018) have gained credibility in image captioning as well, and a comprehensive evaluation of all techniques would serve as a valuable metric against which to compare our model.

II. Methods

In order to develop a comprehensive understanding of model behavior, we try four different models, mixing whether dropout or Bahdanau Attention are used. The models are made up of a combined encoder-decoder architecture. The encoder uses the pre-trained ResNet101 model and a linear layer to change feature dimensions. The decoder model is made up of an embedding layer, a LSTM, and a linear layer. For models with dropout, we run the image features through a RelU function and then a dropout function at a p-value of 0.5. In the decoder model, we run the results of the embedding layer through the dropout function, with a p-value of 0.5. Our linear layers are defined by the following formula:

$$y = xA^T + b$$

Dropout layers are defined sampling from the input with a Bernoulli distribution ($\frac{1}{1-p}$). The embedding layer is a lookup table for all the words in our vocabulary.

We use the Adam Optimizer and Cross Entropy Loss, defined as:

$$\mathcal{L}(x, y) = L = \{l_1, \cdots, l_N\}^T$$

$$l_n = -\sum_{c=1}^{C} w_c log(\frac{exp(x_{n,c})}{\sum_{i=1}^{C} exp(x_{n,i})})$$

We train our model on the Flickr8k dataset for up to 1000 epochs on an NVIDIA GTX 1080 TI GPU 12GB, with a batch size of 50. Then we save model parameters for the lowest loss on the validation set. We split our data three ways: a training set, validation set, and testing set, each set having 30K, 5.2K, and 5.2K data samples, respectively. We use a learning rate of $1e{-}4$ and a weight decay for the optimizer of 0.005. We use an embedding size of 256 and a hidden size of 100. We process all images before feeding them into ResNet by resizing them to 224x224 and normalizing the values per the RGB channel to have a mean of $\{0.485, 0.456, 0.406\}$ and std of $\{0.229, 0.224, 0.225\}$ respectively.

III. Results

The outcome of the pursued investigation display intriguing implications regarding the performance of the models with varying additions. Overall, certain combinations were conducive to higher performance, while others significantly detracted from the encoder-decoder models. An example of a good caption and a bad caption are both shown in Figures 3 and 4, respectively.

Furthermore, the accuracy scores given by the BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, CIDEr, and SPICE metrics for evaluating image captioning algorithms are shown in Table 1. Each metric determines how closely a candidate sentence matches the reference sentence with slightly different evaluation mechanisms, and they are the industry standards for assessing the accuracy of an image captioning model. Accuracy scores are organized such that they are

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
ResNet101	23.50	10.98	4.37	1.71	8.03	21.77	18.77	9.31
ResNet101 - dropout	12.54	5.37	2.12	0.85	7.31	17.08	10.21	8.78
ResNet101 + Attention	13.92	5.28	1.91	0.79	7.01	17.17	6.54	6.79
ResNet101 + Attention - dropout	16.62	6.35	2.22	0.91	7.05	17.78	10.55	7.38

TABLE I

RESULTS OF EACH ITERATION OF RESNET101 IN THIS STUDY. FOR EACH MODEL, WHICH SHOWS WHETHER OR NOT THE DROPOUT FUNCTION AND BAHDANAU ATTENTION ARE INCLUDED, THERE ARE EIGHT EVALUATIVE METRICS THAT ARE THE STANDARD FOR ASSESSING TEXT GENERATION MODELS.

Fig. 3. An example of good captioning. The predicted caption is: "a man in a red shirt is standing on a." The true caption should be: "a man wearing a grey jacket standing on the side of a street."

Fig. 4. An example of bad captioning. The predicted caption is: "a man in a red shirt is standing on a." The true caption should be: "Water streaming from a young woman in a swimming pool flipping her wet hair backward."

shown for ResNet101, ResNet101 with the dropout function, ResNet101 with Bahdanau attention, and ResNet101 with both the dropout function and Bahdanau attention. A closer look at Table 1 illuminates that the performance of the model performs the best on average with no additional Attention or the dropout function. Second to this is ResNet101 with Attention and the dropout function, followed by ResNet101 with just Attention, and lastly by ResNet101 with just the dropout function. Though the score order varies slightly based on the metric used, these results are reported based on an assessment of the average across all eight metrics.

IV. CONCLUSION

The results discussed in Section 3 lend themselves to certain important conclusions, as well as a thorough understanding of the impact and limitations of our study.

A. Impact

Our motivation was partially fulfilled by our results. The ResNet model with dropout can sometimes capture a simple idea of objects in the model, but it fails to capture any specific details about objects. Other times, the model completely fails to capture the relationship between any objects in the image and captions the image incorrectly.

The ResNet with dropout model tends towards writing a more generalized caption that loosely fits a large number of images but does little to illustrate the nuance and details within the image. For example, the ResNet with dropout model predicts all of the images in Figure 3 with caption "a man in a red shirt is standing on a ¡UNK¿ ." For all of the images in Figure 4, the model captions the images "a dog is running in the water ." Clearly, the model can distinguish between an image with a person in it and an image with a dog in it, but it misses details such as the number of objects, gender, and actions. In fact, these two captions are the only two captions that ResNet with dropout predicts for the entire data set.

The ResNet without dropout model makes an attempt at more specific captions for images, but ends up writing captions that not only do not match the images but also do not make sense. For example, "a man in a red shirt is in a red and white and white and white and white and white and white and white and white and white and" is the prediction for the image captioned "Two men , one with purple hair , at crowded fountain near Seattle Space Needle ." While the

Fig. 5. 3 images captioned "a man in a red shirt is standing on a." by ResNet with dropout. Correct captions from left to right: "Some people stand by a group of 5 orange portable bathrooms .", "Person riding their bicycle on the street with a backpack on .", "A woman with dirty blonde hair and sunglasses and a man with dark hair stand in front of a record store ."

Fig. 6. 3 images captioned "a dog is running in the water ." by ResNet with dropout. Correct captions from left to right: "Brown dog chews on bone while laying on the rug .", "a small dog with a blue color fetching a yellow ball", "Two dogs tug at the same item while wearing training gear ."

model has a higher total number of unique predicted captions, none of them do well to caption the images and instead cause the model to write captions that do not capture any idea of the objects in the image, and often do not make logical sense either.

Both models that use Bahdanau attention (dropout and without dropout) attempt to create a generalized caption for the entire dataset. Using only one prediction caption for the entire dataset results in very poor performance, and is not an effective captioning scheme. We would expect that adding Bahdanau attention should improve the performance of our models, rather than limiting them to one prediction for the entire dataset.

Our models captioning ability would have little use in the real world applications that motivated our paper. Image captioning has benefits only while the captions produced are both accurate and specific. To assist visually impaired individuals, a model would need to provide sufficient details to the individual such that they can make out more than just the type of objects in an image. In the medical field, details are also important. A useful model would need to be able to pick up on different shapes, colors, and locations in an image, rather than a generalized observation like identifying an arm.

B. Limitations

Our models had limited success in accurately captioning images. We believe our model may have been limited by the use of ResNet101, as Srinivasan et al. was able to achieve much better captioning scores with a similar model using VGG. We did not do hyperparameter tuning on our models, which may contribute to the inaccuracy of captions. Our choices were decided with some thought but tuning might improve the performance of our models.

Another limitation of our model is the failure of Bahdanau attention. Our models' performance declined when Bahdanau attention was added, when we would expect the opposite. Not being able to utilize attention limited the results of our models.

REFERENCES

[1] Roy, A. (2020, December 9). A Guide to Image Captioning. Medium. Retrieved December 9, 2022, from https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350.

[2] Srinivasan, L., Sreekanthan, D., A.L, A. (n.d.). Image Captioning - A Deep Learning Approach. International Journal of Applied Engineering Research, 13(9).

[3] Timo Ojala, Matti Pietikäinen, and Topi Mäenpä. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In European Conference on Computer Vision. Springer, 404–420.

[4] David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 2 (2004), 91–110.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR'05), Vol. 1. IEEE, 886–893.

[6] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR'05), Vol. 1. IEEE, 886–893.

[7] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 51, 6, Article 118 (November 2019).

[8] Al-Malla, M.A., Jafar, A. and Ghneim, N. 2022. Image captioning model using attention and object features to mimic human image understanding. J Big Data 9, 20 (2022).

[9] Khan, M., Shifath, S.M., and Islam, M.S. (2021). Improved Bengali Image Captioning via deep convolutional neural network based encoder-decoder model. ArXiv, abs/2102.07192.

[10] V. Atliha and D. Šešok, "Comparison of VGG and ResNet used as Encoders for Image Captioning," 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2020, pp. 1-4.

[11] Y. Bhatia, A. Bajpayee, D. Raghuvanshi and H. Mittal, "Image Captioning using Google's Inception-resnet-v2 and Recurrent Neural Network," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-6.

[12] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[13] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR, abs/1409.0473.

[14] Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980.