CS 4782 Final Report

Sunny Sun, Michael Wei, Tony Chen, Linda Hu, Jiye Baek

May 9, 2025

1 Introduction

Image captioning poses the unique challenge of translating visual information into coherent natural language – a task that requires understanding both what is in an image and how those elements relate to one another. In this project, we aimed to reproduce the image captioning results of the paper *Show, Attend, and Tell* published in 2016 by Kelvin Xu *et al.* This paper utilizes the then novel technology of attention to create more accurate and efficient methods of automatic caption generation of a given input image. This paper was motivated by the computer vision problem of determining not only what objects exist within an image, but also accurately describing the relationships between them in a natural language.

Although prior image-captioning networks existed at the time this paper was written, most used static global image features which often miss fine-grained spatial relationships crucial for accurate captioning. This paper's main contributions include introducing two attention-based image caption generators under a common framework, visualizing the results of this framework by outputting alpha masks that depict the area of an image that the attention focused on, and quantitatively validating the usefulness of attention in caption generation using standard metrics of BLEU and METEOR scores. Overall, this paper addressed issues of spatial relationships and focus on specific regions of interest for complex scenes and multiple objects. It is able to output interpretable attention maps and improve caption accuracy and fluency by grounding words in localized visual evidence.

2 Chosen Result

Besides implementing the underlying infrastructure of the model outlined in the paper, we aimed to reproduce two main results:

Visualization of soft attention at each timestep of a generated caption. The original paper has several examples of the change in attention over time to reflect the relevant parts of the image. We chose this visualization to understand the attention mechanism used by the model and gain intuition into what the model saw.

Figure 1: Visual examples of soft attention from the original paper

BLEU and METEOR scores reported for the soft attention model on the Flickr8k dataset. These are common metrics used in natural language processing that quantitatively evaluate the quality of generated captions against human understanding. We chose these results to provide a concrete standard from which we can measure the accuracy of our model.

3 Methodology

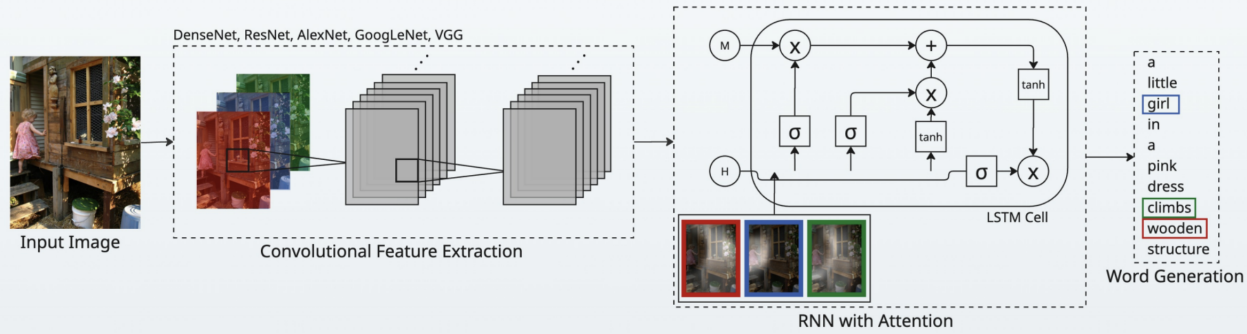

Figure 2: Overall Model Architecture

The model takes in a single raw image and passes it through a convolution neural network (CNN) encoder such as VGG, ResNet, or DenseNet to extract spatial image features in annotation vectors. Instead of training from scratch, we used frozen, pre-trained CNNs encoders to experiment with newer architectures and save compute. We used the encoders from the original paper — VGG, GoogLeNet, and AlexNet — for consistency and baseline comparison. To evaluate the impact of more recent architectures, we also included ResNet and DenseNet, which were introduced after the paper was published. ResNet, with its skip connections, addresses the vanishing gradient problem and enables stable training of deeper networks. DenseNet uses dense connections to promote feature reuse and improve generalization. These design improvements allow both models to capture complex patterns, which likely contributed to their stronger performance, even on our smaller dataset.

The attention model calculates the weights for each annotation vector. These are used in attention mechanism to generate a context vector, indicating where in the image the model attends to at that timestamp. The paper uses two types of attention: hard and soft attention. Soft attention calculates the expected value of all annotation vectors using the softmax of the weights, whereas hard attention uses the weights for each annotation vector to create a multinoulli distribution which is then sampled. Then, hard attention samples from each location of the image as one-hot variable–1 if the i-th location of the image is used to extract visual features, else 0. Thus, the context vector in the hard attention implementation is a multinoulli distribution. We ultimately chose to omit the implementation of hard attention due to its reliance on reinforcement learning, which introduces additional complexity and instability during training.

In the LSTM decoder, we initialized the memory state and hidden states by taking the average of the annotation vectors fed through 2 MLPs. The context vector, previous word, and previous hidden state are fed into LSTM decoder at every word predicted in the caption.

The original paper trained the soft attention model on MS COCO for 3 days on an NVIDIA Titan Black GPU. We trained our re-implemented model over 20 epochs using Adam optimizer with a learning rate of 0.0001 and a batch size of 32. Due to time and memory constraints, we only trained our model on the Flickr8k dataset alone, avoiding Flickr30k and MS COCO in the paper. The smaller dataset also allowed us to maintain fast experimentation cycles and train our model in around 2 hours. We used the train, val, test labels in the dataset to split our data. This gave us 6000 images for the training set and 1000 images each for the validation and test set. In the dataset, there are 5 human annotated captions per image. To evaluate our caption quality, we used BLEU and METEOR scores to calculate the training and validation accuracy. We calculated these scores for soft attention alone, but across all 5 pretrained CNN encoders.

4 Results and Analysis

In terms of BLEU and METEOR scores, our re-implementation achieved slightly lower compared to the original paper's soft attention scores for the Flickr8k dataset. We also trained the model with various frozen,

pre-trained encoders and found that ResNet performed the best and significantly better than AlexNet, especially in terms of BLEU-3 score. As predicted above, ResNet and DenseNet performed the best out of the 5 pre-trained encoders we tested and outputted scores closest to the original paper's performance, being no more than 4% less than any of the original paper's BLEU or METEOR scores. We attribute their high performance to the residual connections which ameliorates the vanishing gradient problem and allows for deeper networks. To account for the discrepancy between our re-implementation and the original paper's performance, we suspect that because the original paper fine-tuned its encoders during training, but our re-implementation used pre-trained, frozen networks. Thus, our encoders do not output representations specific to the task of image captioning as the original paper's encoder does. The scores of our models and the original paper are shown in the table below.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Soft-Attention (Original Paper)	67.0	44.8	29.9	19.5	18.93
AlexNet	60.44	34.90	19.71	10.81	12.89
DenseNet	64.81	40.82	26.26	15.84	16.95
GoogLeNet	64.53	40.13	25.12	14.77	16.45
VGG	64.79	40.37	25.88	15.48	17.08
ResNet	**65.67**	**41.40**	**27.13**	**16.48**	**17.94**

In terms of visualization and qualitative performance, we fed several random images to our model and obtained captions and attention maps after training. Figure 3 demonstrates an example of one such caption and set of attention maps. We observe in the second and third image, the attention maps properly highlight the "red" of the jeep and the "jeep" itself, demonstrating its strength in localization. Additionally, we can see in words such as "is" and "driving" that the model is able to attend to salient regions for words that are non-object. Most importantly, for the word "on", the model strongly highlights the rocky road around the jeep, illustrating its ability to capture the fine-grained relationship between objects in the image–a significant solution to prior image-captioning models. Overall, the model learns alignments that correspond pretty strongly to human intuition, similar to the original paper's findings. Based on the generated attention masks, we would claim that our re-implemented model is capable of building a coherent visual-linguistic representation, grounding each word in relevant spatial context.

Figure 3: Example of our model's attention mechanism visualization per word in generated caption.

One challenge that we faced in the re-implementation process was long training times and timing out on compute while using a T4 GPU on Google Colab. On average, training 10 epochs would take almost two hours. We partially resolved these problems by optimizing the decoder of the model. We also implemented saving model checkpoints at every epoch so we could break up training time over multiple sessions.

5 Reflections

Our re-implementation demonstrated meaningful results given limited resources. Despite only training on the smaller Flickr8k dataset and using frozen pre-trained CNN encoders, our model generated reasonable captions and produced interpretable attention maps. This highlights the effectiveness of the soft attention mechanism, even under constraints.

One of the key lessons we learned was the importance of feasibility given time constraints and tight resources. Although the original paper explored both hard and soft attention mechanisms, we found that soft attention was significantly more practical to implement and train due to its differentiable and deterministic nature. Implementing the hard attention, especially the required reinforcement learning techniques like REINFORCE, from scratch was more time consuming than anticipated.

Overall, this project deepened our understanding of attention mechanisms in image captioning and provided practical experience in re-implementing a significant paper in computer vision and natural language processing. For potential future work and with more time and computational capacity, we would look into using larger datasets that the original paper used such as Flickr30k and MS COCO to improve generalization and caption quality. We can also try to implement hard attention mechanism using REINFORCE to explore sharper spatial focus. Finally, we can fine-tune CNN encoder during training to allow the model to learn image features more tailored to caption generation.

6 References

[TODO DO WE NEED TO CREDIT MORE TOOLS ??]

- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Paper Website. https://kelvinxu.github.io/projects/capgen.html
- Xu, K. Arctic Captions: Theano Implementation of Show, Attend and Tell. GitHub repository. https://github.com/kelvinxu/arctic-captions
- Wong, A.. Show, Attend and Tell: PyTorch Implementation. GitHub repository. https://github.com/AaronCCWong/Show-Attend-and-Tell
- Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artificial Intelligence Research, 47, 853–899. https://doi.org/10.1613/jair.3994
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 32. https://arxiv.org/pdf/1912.01703