From Recurrent Neural Networks to Transformers: Comparing AI Models for Image Captioning

William Spence
Queen's University
20wds@queensu.ca

Thomas Tesselaar
Queen's University
thomas.tesselaar@queensu.ca

Zain Parihar
Queen's University
21zp16@queensu.ca

Ryan Robinson
Queen's University
20rjr3@queensu.ca

Hari Nair
Queen's University
21hpn2@queensu.ca

Abstract—**In this paper, we conduct a comprehensive comparison of Recurrent Neural Networks (RNNs), Gated Recurrent Unit (GRU) networks, Long Short-Term Memory (LSTM) networks, and Transformer models in the context of image captioning tasks using the Flickr8k dataset. Our study aims to systematically evaluate the performance of these architectures under uniform conditions to discern their capabilities and limitations in generating contextually relevant and accurate image captions. We assess each model based on cross-entropy loss and BLEU scores, examining their learning efficiency, susceptibility to overfitting, and overall caption quality. Our results reveal an inverse relationship between model complexity and output quality for our specific dataset and encoder-decoder architecture, highlighting the importance of task-specific model selection.**

I. Introduction

A. Motivation

Image caption generation is an integral challenge within deep learning, situated at the crossroads of computer vision and natural language processing. Image captioning models have a plethora of applications across several fields, especially assistive technologies for the visually impaired (Ghandi et al., 2023). Throughout the evolution of deep learning, a variety of architectures have been proposed to provide solutions, each with their own advantages and drawbacks.

Prior to the advent of the Transformer architecture, state-of-the-art image captioning models often utilized a convolutional neural network (CNN) and recurrent neural networks (RNN) in an encoder-decoder framework, using the CNN for image feature extraction and the RNN for caption generation.

RNN variants, such as Gated Recurrent Unit (GRU) networks and Long Short-Term Memory (LSTM) networks, introduce additional parameters within each unit to provide an internal ”gate” which can be trained to capture long-term dependencies within a sequence. Although these additional parameters better equip the network for training on complex datasets, they can hinder training speed and performance on smaller datasets (Gruber and Jockisch, 2020).

In 2017, the Transformer architecture followed the development of CNN-RNN models, proposed by Google researchers in the groundbreaking paper ”Attention is All You Need” (Vaswani et al., 2017). Unlike its predecessors that rely on sequential data processing, Transformers employ a self-attention mechanism, enabling the model to process all parts of the input data in parallel. This feature not only improves the model's efficiency in learning long-range dependencies within the data but also enhances its capability to generate more contextually relevant captions by correlating specific image features with appropriate textual descriptors. Despite this, Transformers are difficult to implement due to high computational intensity, mostly attributed to the self-attention mechanism's quadratic complexity, which poses challenges for memory consumption and processing power requirements (Vaswani et al., 2017).

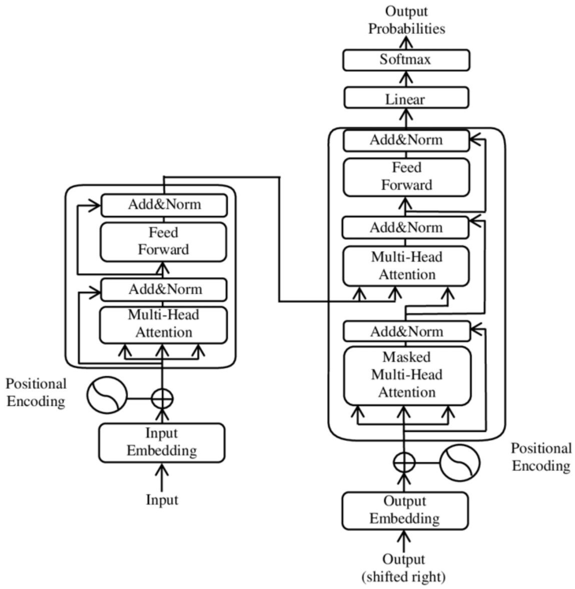

Fig. 1: Visual model of the Transformer Architecture (Vaswani et al., 2017)

B. Problem Definition

Despite thorough evaluations of image captioning methods, comparative research examining these architectures under similar conditions remains scarce. This work aims to bridge this gap by systematically comparing a variety of image captioning architectures under uniform testing conditions. By carefully controlling the experimental environment, a more nuanced understanding of how these models perform relative to each other will be provided.

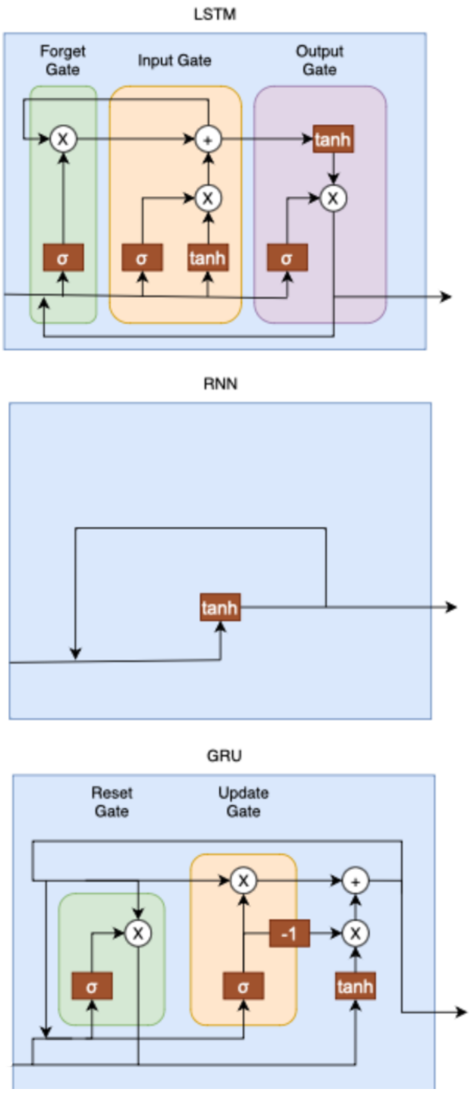

Fig. 2: Visual models of the LSTM, RNN, and GRU Architectures

II. Related Work

A. Show and Tell: A Neural Image Caption Generator

The ”Neural Image Caption” (NIC) model, introduced in the paper ”Show and Tell: A Neural Image Caption Generator”, marked a significant milestone within the domain of image captioning. Prior image captioning solutions combined a variety of existing techniques to detect objects in images and convert them to natural language using rule-based systems (Vinyals et al., 2015). The NIC model was the first solution which used a single joint model to caption images, taking inspiration from machine translation methods. In 2014, machine translation models achieved state-of-the-art performance using an ”encoder” RNN to process input sentences, providing a rich vector representation which is passed into a ”decoder” RNN as the initial state. This solution was mapped to the challenge of image captioning, as the task is essentially ”translating” an image to text.

To provide a rich representation of the input image, a highly-performant CNN pre-trained for image classification was employed, passing its final layer as the initial state for a ”decoder” RNN. The NIC model was trained and tested on four datasets consisting of image-caption pairs, surpassing state-of-the-art BLEU scores on all four datasets. The NIC model revolutionized the domain of image captioning by providing an elegant framework for future models.

B. Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention

The introduction of the ”Show, Attend, and Tell” model further refined neural image captioning, building upon the foundational work established by the ”Neural Image Caption” (NIC) model (Xu et al., 2016). This model, detailed in the paper ”Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention,” represented a pivotal advancement by integrating a visual attention mechanism into the image captioning process. This enhancement allowed the model to generate captions by dynamically focusing on different parts of the image while generating each word of the caption. The key idea was to develop a model which worked by mimicking the human visual attention mechanism.

This model used a similar encoder-decoder framework as its predecessor, the NIC model. However, instead of using a single, static representation of the image as input to the decoder, it introduced an attention mechanism that enabled the decoder to selectively focus on different regions of the image at each step of the caption generation. This approach was inspired by advancements in machine translation, where attention mechanisms had been used to focus on different parts of the source sentence when translating to the target language.

To implement this, the model utilized a CNN as the encoder to extract a set of feature vectors representing different parts of the image. The decoder, an RNN, then generated the caption one word at a time, with the attention mechanism weighing the importance of each feature vector at each step based on the current context and the previous state of the RNN. This process allowed the model to generate more detailed and contextually relevant captions by ”attending” to the parts of the image that were most relevant to each word being generated.

C. BLEU: a Method for Automatic Evaluation of Machine Translation

BLEU, short for Bilingual Evaluation Understudy, is a method initially designed for automatic evaluation of machine translation that has been widely adopted in assessing image captioning models as well (Papineni et al., 2002). The core idea behind BLEU is to compare the machine-generated text to one or more human-written reference texts, focusing on the precision of word sequences that appear in both the generated and reference texts. Essentially, it measures how many words and phrases in the machine-generated captions match those in the reference captions, adjusted for the length of the caption to prevent favoring overly short or nonsensical responses.

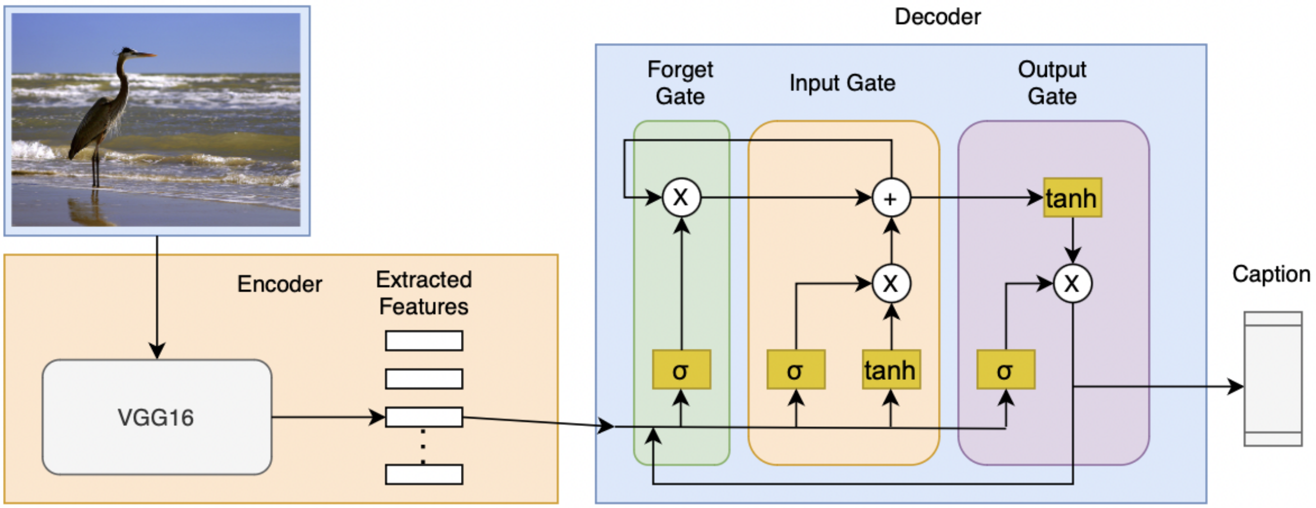

Fig. 3: Depiction of the general encoder-decoder architecture used for image captioning model comparison. Shown with a LSTM as the decoder, which is substituted for a RNN, GRU, and Transformer in our experiment.

To accomplish this, BLEU calculates the n-gram precision for the generated text, which involves counting the matching n-grams (contiguous sequences of n words) in the generated and reference texts and then dividing by the total number of n-grams in the generated text. This is done for various lengths of n-grams (typically up to 4) to ensure that the evaluation captures both the accuracy of individual words and the fluency of longer phrases. BLEU also incorporates a brevity penalty to counteract the bias toward shorter texts; this penalty reduces the score of translations that are significantly shorter than the reference texts, ensuring that the generated captions are not only accurate but also of appropriate length. By combining these elements, BLEU provides a quantitative measure of how closely the generated captions resemble human-written captions, making it a valuable tool for evaluating the performance of image captioning models in producing coherent, contextually appropriate descriptions.

D. Very Deep Convolutional Networks for Large-Scale Image Recognition

”Very Deep Convolutional Networks for Large-Scale Image Recognition” (Simonyan and Zisserman, 2015) was a seminal paper by K. Simonyan and A. Zisserman from the Visual Geometry Group (VGG) at the University of Oxford. The paper was published in 2014, and introduces the VGG network architecture (VGGNet), particularly focusing on VGG16 and VGG19, which are deep CNNs designed for image classification tasks.

The motivation behind these models was to explore the impact of increasing network depth on image classification accuracy. VGG architectures use small (3x3) convolutional filters, and max-pooling layers. The key contribution of the paper was to demonstrate that deeper networks can learn more discriminative features from images, leading to improved classification performance. VGG16 & VGG 19 achieved remarkable results on benchmark datasets, most notably with a strong performance in the 2014 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The paper also demonstrates how the architectures were able to generalise well to other datasets, achieving state-of-the-art results.

The success of VGG16 and VGG19 highlighted the importance of depth in CNN architectures and paved the way for subsequent advancements in deep learning research. The effectiveness and simplicity of the VGG architectures have led to them becoming become widely adopted in various computer vision tasks beyond image classification, including object detection, image segmentation, and feature extraction.

E. Attention Is All You Need

The Transformer model, introduced in the paper ”Attention is All You Need” (Vaswani et al., 2017), revolutionized the field of natural language processing (NLP) and subsequently, various multimodal tasks such as image captioning. Unlike traditional RNNs and LSTM networks, which rely on sequential processing, the Transformer model employs self-attention mechanisms to capture global dependencies within the input sequence efficiently. This mechanism allows the model to attend to all positions in the input sequence simultaneously, enabling parallelization and fixing common issues like vanishing gradients and long-range dependencies encountered in sequential models.

In the context of image captioning, the Transformer model exhibits several performative advantages over RNNs, LSTMs, and other attention-based models. Firstly, its self-attention mechanism enables the model to capture complex relationships between visual features and textual tokens without being constrained by sequential processing, thus facilitating better understanding of long-range dependencies in images. Secondly, the parallelization inherent in the Transformer architecture leads to faster training times compared to sequential models,

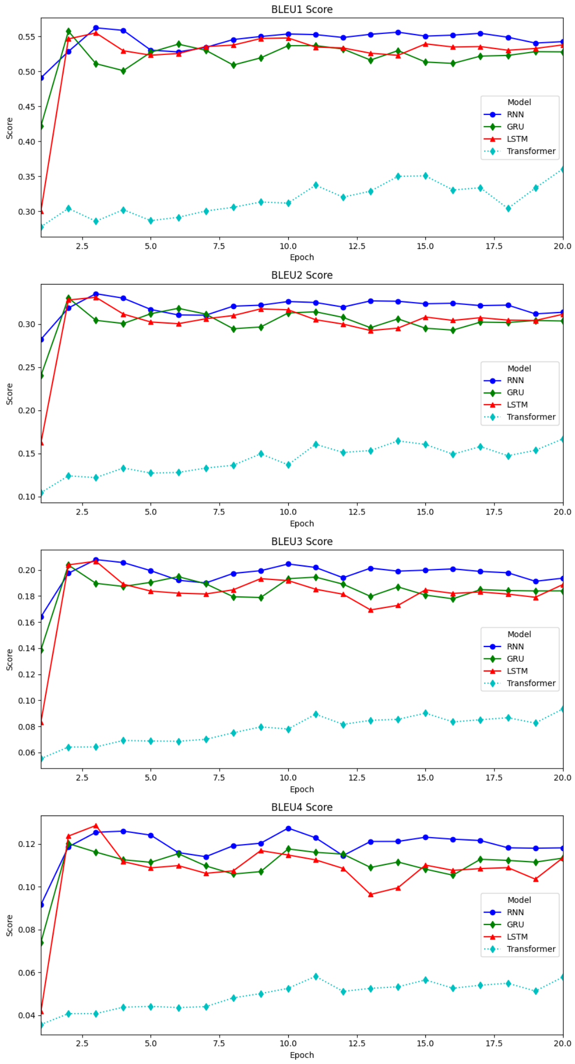

Fig. 4: Plots of each model's BLEU scores for n-grams 1 to 4 versus number of training epochs

making it more scalable for large-scale datasets. However, despite these advantages, the Transformer model may face challenges in handling temporal information and sequential patterns present in images, which have been traditionally well-suited for RNNs and LSTMs. Additionally, the computational complexity of self-attention mechanisms poses scalability issues for extremely large image datasets. Nonetheless, recent advancements in adapting Transformer architectures for image captioning tasks show promising results, underscoring the potential of this model in multimodal learning applications.

III. Methodology

A. Dataset

To accommodate our model training and validation process, we elected to use the Flickr8k dataset. This choice was guided by its manageable size and carefully curated data, ideal for training and evaluating several models. The dataset contains 8000 images, with each image mapped to 5 captions of maximum length 35. In the preprocessing phase, all captions were converted to lowercase, start and end tags were added to signal the start/end of the sequence, and unnecessary spaces at the beginning or end of captions were removed. Finally, all input sentences are tokenized for further processing.

B. Experiment Setup

As stated, the goal of this work is to compare image captioning models under uniform testing conditions. To achieve this, we propose a lightweight encoder-decoder architecture inspired by Google's NIC model (Vinyals et al., 2015). We will be comparing the independent performance of RNNs, GRUs, LSTMs, and Transformers as the decoder section of our models.

For image feature extraction, we employ a pretrained VGG16 CNN model due to its proven ability to generate rich image representations (Simonyan and Zisserman, 2015). The final layer of the VGG16 model, where it produces its image class prediction, is removed so that its output is a 4096-dimensional vector representation of the image. As the pretrained model is already highly performant, it will not be finetuned in the training process to reduce computation. After the image features are extracted, a dropout layer with rate 0.4 is applied to prevent overfitting during training. The remaining image features are then passed to a dense layer with 256 units and a rectified linear unit (ReLU) activation function.

For sequence processing, the tokenized input sentence is padded to the maximum caption length of 35. The processed sequence is then passed through an embedding layer which embeds each word as a 256-dimensional vector. A dropout layer with rate 0.4 is applied before the embedding is passed to the sequence processing layer. This is where the various decoder models are implemented.

The RNN, GRU, and LSTM decoders are each 256 units and use the hyperbolic tangent activation function. A small decoder-only variation of the Transformer model was used, as the work done by a typical Transformer encoder is highly analogous to feature extraction which is handled by the pretrained VGG16 model. The decoder uses cross-attention and self-attention layers to map different image features to word tokens, which are output as the predicted caption.

C. Evaluation Methods

To evaluate the performance of each model, the cross-entropy loss and BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores were calculated between each epoch using a test set of 800 images. The comparison of cross-entropy loss on the test set against the train set after each epoch will illuminate the degree to which each model overfits, while the BLEU scores will measure each model's ability to generate captions which accurately describe the image's content.

IV. Results and Discussion

Our analysis of the four models reveals interesting performance nuances. As can be seen in Figure 4, the output quality of the decoders as measured by BLEU is inversely

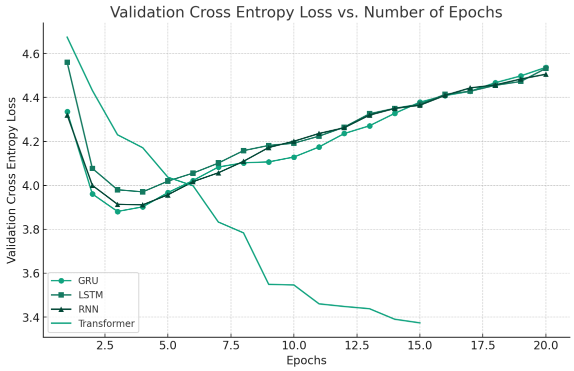

Fig. 5: Plot of each model's cross-entropy loss throughout 20 epochs

proportional to the decoder's complexity, with the simpler models outperforming the models with additional parameters. This expands on the results of Gruber and Jockisch (2020) by demonstrating that RNNs can outperform GRUs, LSTMs, and Transformers in situations with less prevalent data.

Despite efforts to mitigate overfitting through dropout regularization, each of the recurrent models began to show signs of overfitting within the first 5 epochs as seen in Figure 5. Interestingly, the Transformer's performance on the validation set continues to improve each epoch, demonstrating its scalability. Although the Transformer performed significantly worse than the recurrent models, it is clear that additional training would yield better results.

V. Conclusion

This paper investigated the performance of various image captioning models on the Flickr8k dataset. By calculating cross-entropy loss and BLEU scores after each epoch, we demonstrated the speed which RNNs, GRUs, and LSTMs are able to fit to datasets in comparison to Transformers. The experiment also showed that Transformers are less prone to overfitting than the recurrent models, and that their performance will scale proportionately to training time as a result.

The RNN performed better than the other models, despite having the simplest architecture with the fewest parameters. This is likely due to the size of the Flickr8k dataset, which is relatively small compared to datasets which more complex models perform well on. Furthermore, the average caption length in the dataset is 12 words. The gate mechanisms implemented in GRUs and LSTMs are designed to learn long-term dependencies in sequences (Gruber and Jockisch, 2020), but for the datasets short sentences the gates may not be necessary. In this experiment, it appears the additional parameters hinder the model's performance rather than improve it.

To conclude, our results highlight the importance of careful model selection for task specificity. Transformers, GRUs, and LSTMs generally perform better than RNNs in image captioning (Ghandi et al., 2023), but in scenarios with limited data or compute it can be beneficial to select a simpler model. When deciding upon a model, one should consider the limitations of their dataset and computational resources, also collecting empirical data through experimentation to determine which model suits their needs.

VI. Limitations and Future Work

Training and validating four models from zero is a computationally intensive task, hence the Flickr8k dataset was selected for resource efficiency. Had we selected a larger dataset, such as Microsoft COCO or Flickr30k, the more complex models would be able to demonstrate their strengths. Furthermore, increasing the size of the general encoder-decoder architecture with additional hidden layers may produce interesting results, but this was not viable with the compute available to us.

To build upon this work in the future, a more comprehensive analysis should be conducted which evaluates the models on several larger datasets. A variety of hyperparameter configurations should also be examined, evaluating model performances with different loss functions, optimization techniques, quantities of hidden layers, and varying sizes of each layer. Further analysis as described will shed light on the specific circumstances in which each model should be selected.

REFERENCES

Ghandi, T., Pourreza, H., and Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39.

Gruber, N. and Jockisch, A. (2020). Are gru cells more specific and lstm cells more sensitive in motive classification of text? *Frontiers in Artificial Intelligence*, 3.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention.