Efficient Quantized Encoder-Decoder Image Captioning

Masih Zakavi, Jasur Okhunov
sz2803@columbia.edu, jo2662@columbia.edu
April 30, 2024

ABSTRACT

Image Captioning has been a focus of deep learning research since 2015's publication of Show and Tell by Vinyals et al. [6]. The task has come to be viewed as a benchmark for foundation multimodal transformers like Meta's BLIP [4]. However, the massive computational cost associated with training these models presents a significant barrier to those hoping to develop compact image captioning models for specific tasks. In this paper, double quantization is presented as a remedy for the massive computational cost. We train a classic encoder-decoder image captioning model with ViT [9] as encoder and GPT-2 [5] as decoder. Results are presented on the Flickr 8k dataset, showcasing a powerful image captioning model that can be trained quickly and adapted easily for specific applications.

1. INTRODUCTION

The encoder-decoder framework for image captioning was introduced by Vinyals et al. in 2015 [6]. This framework, initially combining an encoder for visual understanding and a decoder for captioning, has paved the way for subsequent advancements. Since the introduction of NLP transformers like the GPT series by Radford et al. [5], the quality of text generation models has continued to improve drastically. Starting with the Vision Transformers (ViT) by Dosovitskiy et al. [9], transformers for vision have continued to grow in popularity as well. These advancements led to the creation of sophisticated and accurate encoder-decoder models for image captioning, such as VinVL by Zhang et al. [8] and Meshed Memory Transformer by Cornia et al. [2]. Today, image captioning models have been largely replaced by multimodal transformers like Flamingo [1] and BLIP [4], which have image captioning as one of their fine-tuning tasks. Yet, the progress in model quality has come at the expense of computational cost. As seen in Figure 1, multimodal transformers have become too large and expensive to be used by those outside major tech firms or universities. This is especially unfortunate given the practical applications in the industry, like surveillance, medical diagnosis, and education, which can benefit from compact image captioning models.

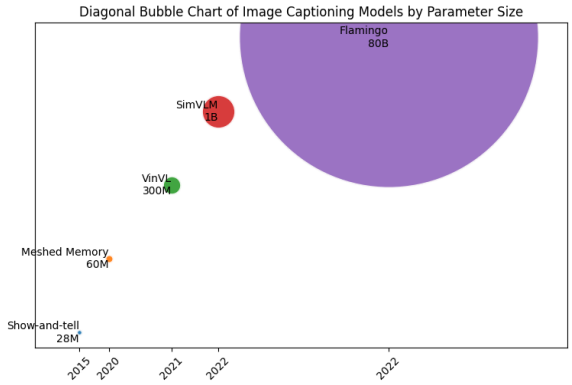

Figure 1: Bubble Chart of state of the art Image Captioning Models by parameter count

An important model compression framework popularized recently is quantization. In fact, the method has proven so effective that recently developed transformer models such as Llama-2 [13] are opting to have 16 floating point data types for their variables or use custom floating point type instead of the traditional 32-bit floating points.

As such, our goal for this project is to develop a compact encoder-decoder image captioning model using quantization. We will specifically quantize both the encoder and decoder and use the Normal Float NF4 data type and double quantization, both introduced in the groundbreaking QLoRA paper. We will apply quantization pre-training and then fine-tune on the quantized model [3].

2. RELATED WORK

2.1. Encoder-Decoder Image Captioning

The inception of encoder-decoder structures in image captioning started with the show-and-tell model [6] that utilized CNNs for image understanding and LSTMs for generating captions. With the rise of transformer-based models, both the complexity and the efficacy of these systems have increased. Vision Transformers have replaced the CNNs, and NLP transformers have replaced LSTMs. However, the encoder-decoder architecture for image captioning has retained its relevance.

2.2. ViT and GPT-2

In our approach, we employ the base version of ViT [9] as the vision encoder and the small variant of GPT-2 [5] as the NLP decoder to maintain a balance between performance and computational efficiency. This choice reflects our goal to demonstrate that our quantization approach is model-agnostic and applicable to a variety of encoder and decoder combinations. The only major requirement is the adaptability of the decoder model, which is the case with most modern LLMs.

2.3. Double Quantization

Since 2016, model quantization has been an active field of study in the deep learning literature. Researchers started by changing model parameters from 32 floating points to 16 floating points with virtually no drop in performance.

Since then, many efforts have been made to compress large deep learning models. An important breakthrough in this field is QLoRA [3], which introduced 4-bit Normal Floats (NF4). This custom data type allows for the optimization of the quantile quantization algorithm. Additionally, [3] introduced double quantization, which involves quantizing the quantization constants to further save on memory requirements and computation time.

For our model, we converted all parameters to NF4 and used double quantization to make the model as compact as possible. One notable detail is that the NF4 algorithm involves transforming model weights for quantization. Since the encoder and decoder have fundamentally different characteristics, we applied the NF4 quantization separately to the two models and then combined the two for training.

3. TRAINING DETAILS

3.1. Dataset

The dataset used was Flickr 8k, which includes over 8,000 images with associated captions. Flickr 8k, together with its larger variant, Flickr 30k, and MS COCO, are the most popular benchmarks for measuring the performance of image captioning models. The Flickr 8k dataset was split 60-20-20 into training-validation-testing.

3.2. Data and Model Processing

Once three data frames with image names and captions were created, we applied the ViT feature extractor to all the images and the GPT-2 [5] tokenizer to all the captions. Once the ViT [9] features were extracted and tokenization was applied to the texts, we created three tensors, one for each training, validation, and testing, and trained the pipeline.

ViT [9] and GPT-2 [5] were both loaded with pre-trained weights, ImageNet classification weights for ViT and language modeling weights for GPT-2. The two models were quantized using double quantization and cast to NF4 separately. Then the two were combined and fine-tuned.

3.3. Training Configurations

All training was done on a single NVIDIA Tesla T4 GPU, which we accessed through Google Colab. A batch size of 16 was used for both training and validation datasets in all results presented. To avoid any loss of knowledge in the pretrained models, a linear warmup schedule was employed. Moreover, gradient decay of 0.01 was used to avoid exploding or vanishing gradients. All models were trained with AdamW optimizer.

GPT-2's remarkable adaptability with minimal fine-tuning was demonstrated in [5]. Since base weights were loaded for both the encoder and the decoder, the fine-tuning required was minimal. As a consequence, we trained the pipeline for a total of 1500 steps (just over 4 epochs).

4. RESULTS

4.1. Metrics

To better understand the results and the quality of our captions, we employed five metrics: ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and METEOR.

BLEU computes the precision of n-grams between machine outputs and reference translations, applying a brevity penalty. BLEU's reliance on n-gram matches limits its ability to recognize semantically equivalent paraphrases.

ROUGE is similar to BLEU, but instead of precision, it measures the recall of n-grams. Its variants, ROUGE-1, ROUGE-2, and ROUGE-L, focus on unigrams, bigrams, and the longest common subsequence, respectively.

METEOR incorporates synonym and paraphrase matching, aligning more closely with human judgment by balancing precision and recall through a harmonic mean. It also adjusts for fluency, considering factors like word order and content word proportions.

Our model achieved a BLEU score of 0.0928, a METEOR score of 0.3908, and a ROUGE-L F1 score of 0.423. Metrics used for text generation models have been criticized in the literature for their failure to align with human judgment, often sacrificing quality for matching words in spite of all the measures set to avoid that. Works like “The Case of BLEU Revisited " by Mathur et al. [12] and TIGERSCORE [11] by Jiang et al. have demonstrated this failure. In light of that, we have provided ten samples in Appendix 1 of varying qualities based on our subjective human judgment. All images were taken from the test samples, and all captions were generated using our optimal model.

4.1 Ablation Study

The natural question is whether the quantized model can match the full model in its caption quality. To answer this question, we conducted ablation studies with a non-quantized pipeline of ViT [9] and GPT-2 [5]. The training hyperparameters used for training the non-quantized model were identical to those used to train the quantized model (A learning rate of 1E-4, linear warm-up schedule with 1000 steps, and weight decay of 0.01). On nearly all metrics, the quantized model outperforms the full model or is on par, as seen in Table 1.

	Full Model	Quant Model
Test Dataset Loss	2.2363	**2.2345**
ROUGE-1 (Precision)	0.3808	**0.4172**
ROUGE-1 (Recall)	**0.498**	0.4966
ROUGE-2 (Precision)	0.1585	**0.1735**
ROUGE-2 (Recall)	**0.2111**	0.2104
ROUGE-L (Precision)	0.3650	**0.4008**
ROUGE-L (Recall)	**0.4779**	0.4775
BLEU Score	0.0690	**0.0928**
METEOR Score	0.3774	**0.3908**

Table 1: Comparison of the full (non-quantized) and the quantized model on caption quality metrics

As mentioned in section 3.3, a major concern for us was the loss of knowledge in the pretrained models in the pipeline. To address this concern, we experimented with different learning rates and warm-up steps. Below are results comparing the performance of models with different learning rates.

Learning Rate	Test Dataset Loss
1e-5	2.5408
5e-5	2.2958
1e-4	**2.2345**
5e-4	2.4271

Table 2: Ablation study results on learning rate

The results and especially the loss curves suggested that once the model is initialized with a lower learning rate, it can converge to a local maxima, and

performance doesn't improve with more training. Similarly, a high learning rate leads to rapid overfitting to the training data with suboptimal generalization.

We also compared different models on their performance with respect to the number of warm-up steps. Results are presented in Table 3.

Warmup steps	Test Dataset Loss
250	2.2932
500	2.3611
1000	**2.2345**

Table 3: Ablation study results on warm-up steps

We found that with a higher number of warm up steps, the model performs better, though unlike with learning rate, we couldn't detect a clear pattern with variation in the number of warm-up steps.

Since the total number of training steps was kept small to limit time and computational resources, we increased the number of steps from 1500 to 4000 to examine whether the results further improved. What we found was that regardless of hyperparameters, the model quickly overfits once the number of steps is increased.

5. CONCLUSION

5.1. Further Work

As Deep Learning models have continued to grow in size, so has the computational resources required to adapt them for specific tasks. While the BFloat16 type has grown more popular over the past 2 years, the framework of quantization remains under-utilized. We hope this paper has been a proof-of-concept for the use of double quantization for Image Captioning and other Encoder-Decoder models. A natural extension of this work would be to experiment with larger LLMs with stronger few-shot performance than GPT-2.

5.2. Acknowledgements

We also relied on the tutorial notebooks by the HuggingFace staff on model compression and EncoderDecoder classes. Without detailed instructions, the training would not have been possible. For this, we are deeply grateful.

5.3. Contributions statement

The conceptual aspects of the model were conceived jointly by the two authors. Coding was mostly done by M.Z., while the write-up was mostly completed by J.O. Both authors trained models for the ablation studies.

6. BIBLIOGRAPHY

[1] Alayrac, Jean-Baptiste, Jonathan Uesato, Po-Sen Huang, Albin Cassirer, Serkan Cabi, Arthur Mensch, Michela Paganini, et al. "Flamingo: a Visual Language Model for Few-Shot Learning." 2022. arXiv:2204.14198 [cs.CV].

[2] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." 2020. arXiv:1912.08226 [cs.CV].

[3] Kim, Woncheol, Dong-Jin Kim, Jongseok Kim, Gunhee Kim, and Yunjae Jung. "QLoRA: Query-focused Low-rank Adaptation for Text-to-image Generation." 2022. arXiv:2205.15839 [cs.CV].

[4] Li, Jiahui, Ning Ding, Zhe Gan, Haofan Wang, Zicheng Liu, Jianfeng Gao, Yumao Lu, and Yalda Uhls. "BLIP: Bootstrapped Language Image Pre-training for Vision-Language Foundation Models." 2022. arXiv:2201.12086 [cs.CL].

[5] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners." 2019. OpenAI Blog.

[6] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and Tell: A Neural Image Caption Generator." 2015. arXiv:1411.4555 [cs.LG].

[1] Alayrac, Jean-Baptiste, Jonathan Uesato, Po-Sen Huang, Albin Cassirer, Serkan Cabi, Arthur Mensch, Michela Paganini, et al. "Flamingo: a Visual

Language Model for Few-Shot Learning." 2022. arXiv:2204.14198 [cs.CV].

[2] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." 2020. arXiv:1912.08226 [cs.CV].

[3] Kim, Woncheol, Dong-Jin Kim, Jongseok Kim, Gunhee Kim, and Yunjae Jung. "QLoRA: Query-focused Low-rank Adaptation for Text-to-image Generation." 2022. arXiv:2205.15839 [cs.CV].

[4] Li, Jiahui, Ning Ding, Zhe Gan, Haofan Wang, Zicheng Liu, Jianfeng Gao, Yumao Lu, and Yalda Uhls. "BLIP: Bootstrapped Language Image Pre-training for Vision-Language Foundation Models." 2022. arXiv:2201.12086 [cs.CL].

[5] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners." 2019. OpenAI Blog.

[6] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and Tell: A Neural Image Caption Generator." 2015. arXiv:1411.4555 [cs.LG].

[7] Wang, Weizhen, Ming Yan, Fei Wang, and Chen Wu. "SimVLM: Simple Visual Language Model Pre-training with Weak Supervision." 2021. arXiv:2108.10904 [cs.CV].

[8] Zhang, Pengchuan, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. "VinVL: Revisiting Visual Representations in Vision-Language Models." 2021. arXiv:2101.00529 [cs.CV].

[9] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." 2020. arXiv:2010.11929 [cs.CV].

[10] Lee, Jun Haeng, Sangwon Ha, Saerom Choi, Won-Jo Lee, and Seungwon Lee. "Quantization for Rapid Deployment of Deep Neural Networks." 2018. arXiv:1810.05488v1 [cs.NE].

[11] Jiang, Dongfu, Li, Yishan, Zhang, Ge, Huang, Wenhao, Lin, Bill Yuchen, and Chen, Wenhu. "TIGERSCORE: Towards Building Explainable Metric for All Text Generation Tasks." University of Waterloo, Tsinghua University, IN.AI, Allen Institute for AI. 2023. arXiv:2310.00752v3 [cs.CL].

[12] Mathur, Nitika, Baldwin, Timothy, and Cohn, Trevor. "Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics." School of Computing and Information Systems, The University of Melbourne.

[13] Touvron, Hugo, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." GenAI, Meta. 2023. arXiv:2307.09288v2 [cs.CL].

7. APPENDIX

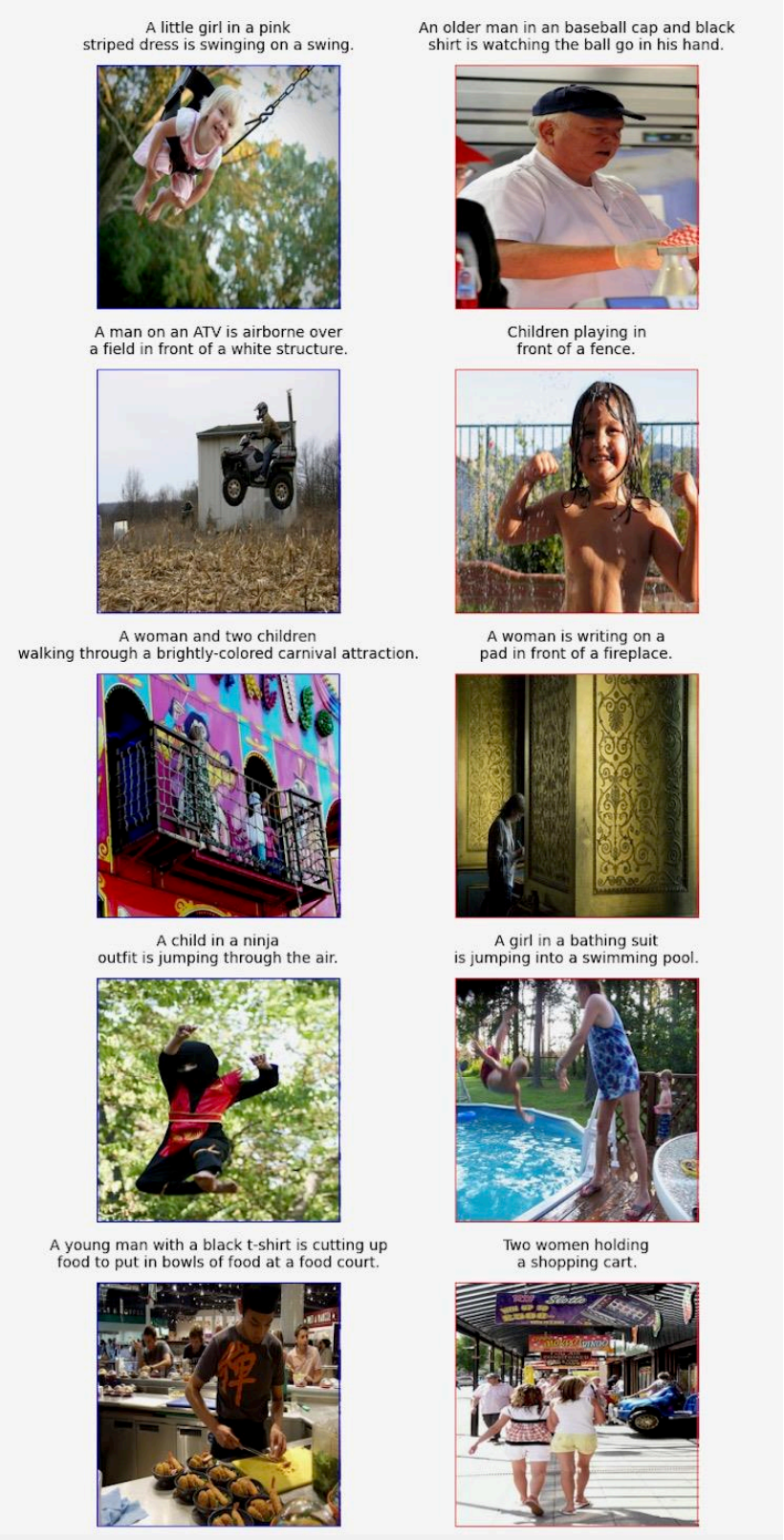

Figure 2: The images in the left column are examples of high-quality captions while those on the right present samples where the model produced weaker captions