SmoothQuant & Activation Aware Weight Quantization for Vision Language Models

Tahmid Jamal
MIT
77 Massachusetts Avenue, Cambridge 02139
tahmid@mit.edu

Zoe Wong
MIT
77 Massachusetts Avenue, Cambridge 02139
zoewong@mit.edu

Rachelle Hu
Wellesley College
21 Wellesley College Road Unit 4925, Wellesley 02481
rh103@wellesley.edu

Abstract

This paper explores the acceleration of vision-language models (VLMs) by leveraging activation-aware weight quantization (AWQ) and SmoothQuant techniques. We focus on quantizing the BLIP-2 architecture, a pre-training method integrating large language models (LLMs) with image encoders. The evaluation is conducted on BLIP-2 models of 2.7B and 6.7B parameters, addressing challenges posed by memory constraints. SmoothQuant, designed to ease the quantization of activations by smoothing outliers, is applied to BLIP-2 models, demonstrating its impact on both the language and image encoder components. AWQ, a method focusing on weight quantization, is employed to achieve significant reductions in model size without compromising performance. Evaluation metrics, including CIDEr and SPICE, are implemented to assess the similarity between generated and reference captions. The results indicate that W8A8 quantization performs as expected, W4A4 per-tensor and per-channel or per-token quantization leads to nonsensical outputs, but W4A4 grouped quantization produces good outputs, mitigating performance degradation.

1 Introduction

Large language models (LLMs) have displayed good understanding of contextual awareness and introducing other modalities such as vision has shown that we can also extract scene understanding as well. Most applications at the time are focused on tasks of single image captioning or question answering. A natural extension is real time scene captioning. Although real time captioning may be difficult, even a dramatic speedup in inference time would still be very useful for scene captioning, especially in settings for the visually impaired. The first reasonable step is then to speedup inference in general for images.

For this project, we worked on implementing smoothQuant and activation aware weight quantization for the purpose of quantizing the LLM of the vision language models (VLMs) since the LLM dominates the entire model size. We replicated the results and then experimented with activation

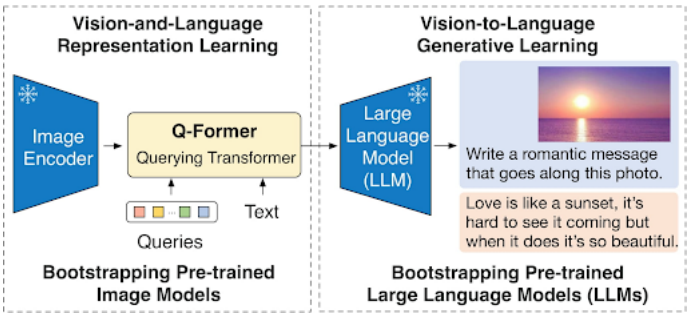

Figure 1: BLIP-2 architecture

quantization as well for 4-bit weight, 4-bit activation (W4A4) quantization. We aimed for the quantized VLM to offer faster and more efficient inference performance, providing accurate and rapid captions or answers to support visually impaired people.

We evaluate these quantization methods on BLIP-2, a pre-training method that integrates frozen pre-trained LLMs and image encoders for vision-language pre-training [2]. More specifically, as shown in 1, BLIP-2 consists of a vision model and language model, connected by a 12-layer Transformer encoder called a Querying Transformer (Q-Former). BLIP-2 is available on HuggingFace, with a suite of models released by Salesforce in varying sizes, both pre-trained and fine-tuned on datasets. Due to memory constraints, we only quantized the BLIP-2 2.7B and 6.7B models.

2 Quantization Methods

We compare previously introduced methods for quantizing LLMs, including SmoothQuant and AWQ, to our proposed method, where we first apply AWQ, then SmoothQuant for W4A4.

2.1 AWQ

Activation-aware Weight Quantization (AWQ) solely quantizes weights but allows for INT3/4 weight quantization without mixed precision, greatly reducing the space needed for LMs while being hardware-friendly [3]. It maintains performance by preserving the most salient weights of a LLM by observing the magnitude of the activations and scaling up the weights that produce the greatest values. The uniform bit size for weights allows for AWQ quantized models to be used across platforms without specialized hardware design. AWQ is additionally effective for both instruction-tuned LMs and even multi-modal LMs, so it is well-suited for BLIP-2.

We select the salient weights by choosing those that produce the top 1% of activations by magnitude, and scale them up. In our experiments, we grid search for α to find the best value to scale up the weights. We do this by optimizing this objective function, where Q is the quantization function, $\mathbf{W}$ are the original weights, $\mathbf{X}$ is the input, and $\mathbf{s}$ is the per-channel scaling factor:

$$\mathbf{s}^* = \text{argmin}_{\mathbf{s}} L(\mathbf{s}), L(\mathbf{s}) = \|Q(\mathbf{W} \cdot \mathbf{s})(\mathbf{s}^{-1} \cdot \mathbf{X}) - \mathbf{W}\mathbf{X}\|$$

Once $\mathbf{s}^*$ scalars are chosen, the weight matrix $\mathbf{W}$ is scaled and the division is baked into a a layer normalization. This led to an improvement in the model output, reducing hallucination and maintaining the captions' similarity to the original FP16 model.

2.2 SmoothQuant

Quantizing activations is difficult because the majority of activations are low effective bits with a few high-magnitude outliers. This is especially the case as model size increases beyond 7B parameters. SmoothQuant aims to migrate the difficulty of quantization from activations to weights by smoothing the activation outliers [5]. In the paper that introduced this method, the evaluation of SmoothQuant was performed on a suite of LLMs including OPT, BLOOM, GLM, MT-NLG, and LLaMA to enable INT8 (W8A8) quantization.

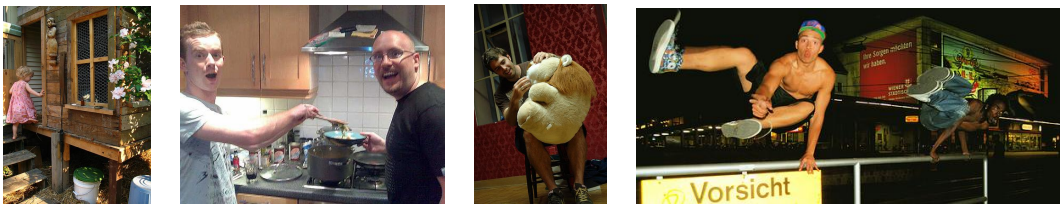

Figure 2: Images sampled from the Flickr30k dataset, which we refer to as A, B, C, and D in following figures from left to right, respectively.

SmoothQuant scales both activations and weights while maintaining the mathematical equivalence:

$$\mathbf{Y} = (\mathbf{X}\text{diag}(\mathbf{s})^{-1}) \cdot (\text{diag}(\mathbf{s})\mathbf{W}) = \hat{\mathbf{X}}\hat{\mathbf{W}}$$

Here, $\mathbf{X}$ is the original input, $\mathbf{W}$ are the original weights, and $\mathbf{s}$ is the per-channel scaling factor. $\hat{X}$ and $\hat{W}$ are the resulting input and weights. We use the loss optimization function described for AWQ to find the scaling factor for SmoothQuant.

We apply SmoothQuant on the BLIP-2 models described in the Introduction and compare its performance on BLIP-2 2.7B & 6.7B when quantizing per-tensor or per-channel and per-token. As a baseline to compare to our method, we showed that W8A8 quantization maintains the model accuracy and also applied SmoothQuant for W4A4 quantization with less performant results. Furthermore, the original SmoothQuant approach only quantizes the LM component of BLIP-2, so we additionally quantize the image encoder component of BLIP-2 in our application of SmoothQuant.

Another important aspect of our approach was the granularity of the quantization. Per-tensor quantization uses the same scaling factor for a tensor, but we also experimented with per-channel quantization for weights, per-token quantization for activations, and grouped quantization, allowing us to achieve higher accuracy despite lowering bit size.

2.3 AWQ & SmoothQuant

SmoothQuant by itself does not achieve performant W4A4 quantization on multi-modal LMs, and AWQ does not quantize activations. We proposed first generating scaling factors based on the AWQ approach with a group-aware optimization function, then using the scaling with SmoothQuant to enable W4A4 quantization.

3 Evaluation Metrics

To evaluate and compare the performances of the models produced from the various quantization methods, we implemented the following metrics: CIDEr and SPICE. Both are measures of how similar the generated caption and the reference captions are. These metrics are evaluated on the Flickr30k dataset available on HuggingFace [6].

We installed and utilized the `pycocoevalcap` python library to generate both scores.

3.1 CIDEr

The CIDEr, or Consensus-based Image Description Evaluation, metric measures the similarity between generated and reference captions [4]. A higher CIDEr score typically indicates a better resemblance between the new generated caption to the references ones.

CIDEr attempts to determine human consensus or sentence similarity based on n-gram overlap. A n-gram is essentially a subsequence of n words in a given sentence, where $n \in [1, 4]$. Each of the generated and reference sentences are represented as a set of n-grams present in it, and each n-gram is given a weight according to its Term Frequency Inverse Document Frequency (TF-IDF). TF places a higher weight on n-grams that appear frequently in the sentences, while IDF gives a higher weight to n-grams that rarely appear in the entire caption dataset. Intuitively, together TF-IDF is a measure of how important a n-gram is within its sentence. The CIDEr score for n-grams of a given length n

A. a child in a pink dress is climbing up a set of stairs in an entry way
B. two men, one in a gray shirt, one in a black shirt, standing near a stove
C. a man sits in a chair while holding a large stuffed animal of a lion
D. two men in germany jumping over a rail at the same time without shirts

Figure 3: Samples of original captions

A. a little girl is climbing up a ladder to get into a house
B. two men in a kitchen preparing food in a pot on the stove
C. a man sitting on a chair holding a stuffed animal lion
D. two men are doing tricks on a railing in the city

Figure 4: Samples of original model outputs

is defined as the average cosine similarity between the candidate and reference sentences. The overall CIDEr score is computed by averaging across all values of n from 1 to 4.

3.2 SPICE

The introduction of SPICE, or Semantic Propositional Image Caption Evaluation, attempts to solve one of the problems with using CIDEr, and that is that n-gram overlap is not necessarily important for two sentences to have the same meaning [1]. Instead, in the computation of the SPICE score, both the generated and reference captions are transformed into a scene graph, or essentially a graph-based semantic representation that records only the objects, attributes, and relationships, filtering out most of the unique language nuances and syntax of the natural language. Thus, the SPICE score is then a measure of the similarity of the candidate and reference scene graphs.

Similar to the CIDEr score, a higher SPICE score indicates a greater semantic similarity between the generated and reference captions.

4 Results

SmoothQuant			CIDEr	SPICE
BLIP-2 6.7B	Original		0.773	0.172
	W4A4	tensor	0.0	0.0
		grouped ($q = 128$)	0.725	0.169
	Original		0.799	0.175
	W8A8	tensor	0.699	0.158
	W4A4	tensor	0.0	0.0
		grouped ($q = 128$)	0.673	0.156

Table 1: Performance after applying SmoothQuant on BLIP-2

We successfully apply W4A4 quantization on BLIP-2 and achieve sensical outputs. Figure 2 is a sample of four images that we display the model outputs for. Figure 3 shows the ground-truth captions given in the Flickr30k dataset, figure 4 includes the original (not quantized) model outputs, figure 5 displays outputs from the W8A8 quantized model, and figure 6 shows W4A4 group-quantized model

A. a little girl standing on a wooden porch with steps leading up to it
B. two men standing in a kitchen with a pot of food
C. a man sitting on a chair holding a stuffed animal lion
D. two men doing tricks on a skateboard at night on a street

Figure 5: Samples of BLIP-2 6.7B W8A8 per-tensor quantization outputs

A. a little girl is climbing up the steps to the house
B. two men in a kitchen preparing food for a meal in a pot
C. a man sitting on a chair holding a stuffed animal lion
D. two men are doing tricks on a railing in the city

Figure 6: Samples of BLIP-2 6.7B W4A4 grouped quantization outputs

AWQ		CIDEr	SPICE
BLIP-2 6.7B	Original	0.799	0.175
	Image encoder & LLM	0.756	0.167
	LLM only	0.797	0.175

Table 2: Performance after applying AWQ on BLIP-2

W4A4		CIDEr	SPICE
BLIP-2 6.7B	Group-Aware SmoothQuant	**0.727**	**0.158**
	Vanilla SmoothQuant	0.673	0.156

Table 3: Performance comparing W4A4 with Vanilla SmoothQuant and Group-Aware SmoothQuant on BLIP-2

outputs. We observe that the baseline BLIP2 model tends to mention skateboards a lot, despite there not necessarily existing one in the image. Meanwhile, the A4W4 BLIP2 model tends to mention skateboards less

Table 1 displays the performance of BLIP-2 after applying SmoothQuant. We compare the 2.7B and 6.7B model sizes, number of bits W8A8 and W4A4, as well as the granularity of quantization. The results show that W8A8 quantization preserves a good amount of the lexical content in captions. W4A4 produced mostly nonsensical content and the performance greatly decreased for both model sizes with per-tensor and per-channel/per-token quantization. However, introducing grouped quantization with group size $q = 128$ gave better captioning and maintained performance.

Table 2 shows the performance of AWQ (W4A16) on BLIP-2 models depending on the components of the model we quantize. We see that, as expected, quantizing the LLM maintains the performance of the model. Additionally, quantizing the image encoder decreases performance but may not lead to a significant improvement in space, as the LLM is generally the largest part of the model.

Table 3 shows our final W4A4 results, comparing the vanilla SmoothQuant approach to Group-Aware SmoothQuant mentioned in Section 2.3. We find that our approach maintains the performance of the BLIP-2 model while achieving lower bit quantization.

5 Conclusion

We reproduced previous SmoothQuant and AWQ quantization methods on BLIP-2, an open-source VLM to achieve performant W8A8 and W4A16 quantized models. We also found that W4A4 is achievable with group quantization for VLMs, despite per-tensor and per-channel/per-token producing nonsensical captions. We would like to apply these methods to larger models, as we were resource-constrained by disk and memory space. As mentioned in the paper, the benefit of the SmoothQuant approach becomes more apparent as activation outliers grow with model size. Lastly, it would be ideal if future work could deploy the quantized model on an edge device, such as a Jetson Nano.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[3] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

[4] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[5] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[6] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.