DI725 - Project Proposal
Quantized Vision-Language Adapter (QVLA) for Efficient Fine-Tuning and Inference

Nisan Yıldız
Dept. of Medical Informatics
Informatics Institute, METU
Ankara, Turkey
nisany@metu.edu.tr

***Abstract*—Recent advances in vision language models (VLMs) favor large base models followed by task-specific fine-tuning. With such a framework, full fine-tuning has become another associated cost for deploying VLMs. To address this cost, parameter efficient methods like Adapters and LoRA have been developed, with QLoRA introducing quantization for higher efficiency. Here, we propose the Quantized Vision Language Adapter (QVLA), based on the VL-Adapter and combining quantization with adapters to enable efficient, scalable fine-tuning. Combined with the PaliGemma base-model, a VLM emphasizing a compact and efficient architecture; QVLA aims to offer practical deployment in resource-constrained settings.**

I. Introduction

Large scale neural networks trained on massive datasets have led to significant breakthroughs in unimodal domains such as natural language processing (NLP) and computer vision (CV). However, many real world problems require inferences across multiple modalities. Vision Language Models (VLMs) have emerged as a key era of research, incorporating visual and textual modalities in a unified framework.

A recent trend in VLM development mirrors those in NLP [1] [2], where pre-trained (large) base models are later fine-tuned for various applications. This paradigm has incentivized ever increasing numbers of parameters, reaching billions in the base models. As a result, it has also become increasingly cumbersome to fine-tune these models with the traditional way of updating the entire set of billions of parameters. In addition to the computational cost of fine tuning, fine-tuning can also make large models susceptible to catastrophic forgetting [3]. This prompted research into parameter efficient fine-tuning (PEFT) methods, where the aim is to limit the number of parameters updated during fine-tuning.

Research on parameter efficient fine-tuning has mainly focused on prompt-tuning or methods that fine-tune over a significantly decreased size of parameters. We will not be going into the details about prompt-tuning here. Examples of the latter approach include Adapter based methods where small adapter heads are inserted into the pre-trained model and the resulting model is fine-tuned with (usually) the pre-trained weights frozen; it also includes methods like LoRA where a low rank representation of the model is trained with the frozen weight of the pre-trained model and concatenated with the original model. Both of the latter approaches perform similarly in terms of training cost and accuracies [4] [5] [6], however, Adapter based methods incur a small inference cost in contrast to LoRA.

For tasks that require a good degree of interaction between different modalities, such as image captioning or visual question answering, Adapter architectures with cross-modal interactions, such as interactions between image encoders and text decoders, have been shown to perform the best, in contrast to architectures that incorporate non-interacting Adapters [5]. One such example is the VL-Adapter [4], which has been shown to offer better accuracy than LoRA in image captioning tasks with similar training cost. Quantization is a method to decrease memory usage and computational costs of neural networks in both pre- and post-training by lowering the bit lengths of parameters saved [7]. Recently, QLoRA introduced a method to quantize the base model to a much lower memory footprint, with higher precision operations for Low Rank Matrix updating; achieving high performance with substantial memory and computational savings [8]. Although QLoRA focuses on LoRA based fine-tuning, its general findings are applicable to other Adapter based fine-tuning methods. However, effects of quantization with adapters have not yet been thoroughly investigated in the VLM field.

II. Project Proposal

Here, we propose the Quantized Vision Language Adapter (QVLA) that incorporates quantization with the VL-Adapter framework to achieve efficient fine-tuning of Vision Language Models. We use PaliGemma [9], a pre-trained VLM model intended to be a versatile base for various fine-tuning applications, as our base model. The PaliGemma family of models propose compact architectures compared to similar performing contemporary models, that seek to maintain performance while improving efficiency.

We propose a 4-bit quantization of the base model weights, followed by a 16-bit data type used during weight update computations, similar to that of QLoRA. This has been shown to significantly decrease the memory cost while having a performance similar to that of a full 16-bit model. We use the VL-Adapter as proposed, with the vanilla Adapter architecture [10]; we also aim to experiment with the number of Adapter heads.

We will compare the following configurations on the RISC image captioning dataset, in terms of accuracy, training and inference times:

- Traditional fine-tuning of the PaliGemma model
- VL-Adapter fine-tuning without quantization
- QVLA (quantization + Adapters)

The RISC dataset includes a total of 44521 satellite images with 5 captions per image that briefly summarize it, with a total of 222605 captions.

III. Code and Data

All of the code associated with the project will be available at: https://github.com/NisanYildiz/DI725-project

References

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[3] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," 2025. [Online]. Available: https://arxiv.org/abs/2308.08747

[4] Y. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," *CoRR*, vol. abs/2112.06825, 2021. [Online]. Available: https://arxiv.org/abs/2112.06825

[5] J. Xing, J. Liu, J. Wang, L. Sun, X. Chen, X. Gu, and Y. Wang, "A survey of efficient fine-tuning methods for vision-language models — prompt and adapter," *Computers Graphics*, vol. 119, p. 103885, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0097849324000128

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[7] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," *CoRR*, vol. abs/2106.08295, 2021. [Online]. Available: https://arxiv.org/abs/2106.08295

[8] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[9] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai, "Paligemma: A versatile 3b vlm for transfer," 2024. [Online]. Available: https://arxiv.org/abs/2407.07726

[10] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2790–2799. [Online]. Available: https://proceedings.mlr.press/v97/houlsby19a.html