Quantizing Image Captioners using QLORA for Medical Images

Sam Briggs
Dept. of Linguistics
University of Washington
briggs3@uw.edu

Iris Zhou
Dept. of Mathematics
University of Washington
irisz1@uw.edu

Abstract

Image captioning lies within the intersection of computer vision (CV) and natural language processing (NLP), with the task being the generation of a caption for a given input image. With the recent rise of transformers and large language models (LLMs), there also came research on LLM compression and quantization of parameter-efficient fine-tuning (PEFT) models for space and computation efficiency. Our contribution is the extension of such state-of-the-art quantization methods of PEFT models to the image captioning domain, something that has historically been understudied. Specifically, in this paper we apply the state-of-the-art quantization method, QLoRA [7], designed primarily for LLMs to the state-of-the-art image captioner BLIP-2 [17]. Specifically, we fine-tune BLIP-2 using QLoRA on ROCO [23], extending BLIP-2 to the medical image captioning domain. To the best of our knowledge, this is the first time that empirical tests have been run where the PEFT model QLoRA is applied to image-captioning done by BLIP-2. Our fine-tuned BLIP-2 model using QLoRA makes leaps and bounds over our baselines, obtaining a CIDEr score which more than doubles the CIDEr score obtained by our baselines.

1. Introduction

Image captioning is a vision-language task in which a model takes in an image and outputs a description of the objects and actions occurring in that image. The added complexity of performing both object detection and description generation makes image detection a memory and compute intensive task.

More broadly, lack of memory for model training is a major constraint for researchers who may not have access to the same computational resources as those at large corporations. To address this issue, Dettmers et al. developed Quantized Low Rank Adaptation (QLoRA) [7], a fine-tuning approach on LLMs that was able to achieve the same level of performance on a 65B parameter model with only 6% of the original >780GB of GPU memory.

Since its release, QLoRA usage has been largely focused on memory-efficient fine-tuning of generative language models. However, to the best of our knowledge, QLoRA has not yet been applied to medical image-captioning. We aim explore QLoRA's capabilities in image captioning by applying QLoRA to BLIP-2, a vision-language model that supports a variety of tasks, including image captioning. Although BLIP-2 takes strides toward computational efficiency by bootstrapping pre-trained vision and language models, freezing these models during pre-training, and converting parameters to FP16 in pre-training, BLIP-2's models for image captioning still use 1.1B trainable parameters, making it computationally intensive to train.

We aim to extend BLIP-2's capabilities by using applying QLoRA to BLIP-2 and further fine-tuning on an medical imaging dataset, Radiology Objects in COntext (ROCO) [23].

Since BLIP-2 was pre-trained on natural image data from datasets like COCO [20], Visual Genome [15], and LAION400M [26], we aim to further fine-tune BLIP-2 on ROCO to extend its capabilities for medical image captioning. The typical image found in COCO, like a parked car, and the typical image found in ROCO, like an MRI of unusual neural activity, differ greatly. This contrast increases the difficulty of the image captioning task.

Through this process, we aim to fine-tune a BLIP-2 image captioning model with minimal performance degradation on a single GPU.

The ability to perform memory-efficient fine-tuning on large image captioning models can increase accessibility and lower the barrier to research in image captioning and other vision-language tasks, as well as in the medical domain. The scale of annotated biomedical images is also severely limited in comparison to datasets like ImageNet or COCO. A biomedical image-captioning model could help to produce more annotated data, expanding the potential of language-vision models in the medical domain overall.

2. Related Work

2.1. Image Captioning

The task of image captioning has a rich history. Historically, many traditional statistical and machine learning techniques have been applied to solve this task, but more recently deep learning techniques have become the prominent method for image captioning [11, 28, 30].

Traditional deep learning techniques have involved some sort of image encoder to extract image features, and a language model to generate a caption [11]. Encoder-decoder architectures have been proposed in the past for this task. For example, Vinyals *et al*. [33] proposed using a CNN to encode image features, and an LSTM to decode and generate captions, while Gu *et al*. [9] proposed using a CNN as the language decoder. AlexNet [16] and VGGNet [29] were predominantly used CNN architectures for image encoding [11].

With the advent of attention, and the subsequent rise and prominence of transformers, transformers have been increasingly used for the image captioning task [28]. For instance, Jiang *et al*. [12] used a transformer to help decode image features, reaching state-of-the-art results in 2021. Li *et al*. [18] created BLIP, using multiple transformers to encode image features, as well as generate image captions. More recently, Li *et al*. [17] created BLIP-2. BLIP-2 improves upon BLIP by introducing a transformer between a frozen pre-trained image encoder[1] and a frozen pre-trained large language model (LLM)[2] to achieve state-of-the-art results and speed up computation.

2.2. Quantization

The research space of LLM compression for the purpose of reducing space as well as run time efficiency has a varied literature, with four major methods: (1) *pruning*, where a heuristic is defined to drop redundant parameters, (2) *knowledge distillation* where a larger more robust LLM is used to help transfer its knowledge to a smaller less robust LLM, (3) *low-rank factorization* where model weight matrices are approximated using matrix multiplication of smaller matrices, and (4) *quantization* where floats are converted into discrete forms to take up less memory [37].

Within quantization, there are two major methods: (1) *Post-training Quantization*, where a LLM is trained and then its pararmeters are quantized after the fact (*e.g*. LLM-QAT [21]) and (2) *Quantization Aware Training* where quantization is used throughout the training process [37] (*e.g*. PEQA [14] and QLoRA [7]).

PEQA and QLoRA specifically are quantization methods for Parameter-Efficient Fine-Tuning (PEFT) where a subset of model parameters are trained during the fine-tuning process [14]. PEQA quantizes the weight matrices of a LLM by creating a frozen low-bit weight matrix and a scalar vector, the latter which is fine-tuned for downstream tasks. QLoRA uses a similar fine-tuning strategy as PEQA, but they make improvements by proposing 4-bit Normal Floats (NF4) which perform better than 4-bit floats, double quantization methods where they quantize the quantization constants (scalar vector), and Paged Optimizers where the GPU and CPU seamlessly hand-off computation when the GPU has run out of memory. PEQA [14] showed that PEQA is able to perform comparably with highly compressed LLMs, while [7] was able to show that QLoRA is able replicate state-of-the-art results produced by 16-bit LLMs. For this reason, we use QLORA as our PEFT method.

2.3. Quantization of Image Captioners

As all generative image captioners possess a language model, it feels natural to try to compress such models for computational and space efficiency. Unfortunately this doesn't seem the case, as there are not many attempts to compress image caption models. Two successful attempts include Rampal and Mohanty [25] who compressed a CNN image encoder and LSTM as the language model and acheived comparable scores as their full-unquantized model trained from scratch, and Atlia and Šešok [1] who performed pruning and quantization on the decoder achieving comparable results to their non-compressed image caption models. To the best of our knowledge, QLoRA has never been attempted on BLIP-2.

2.4. Medical Image Captioning

Many previous attempts towards image captioning in the medical domain have been made, with methods corresponding with the more general image captioning domain [3]. For instance, earlier CNN models were proposed by Harzig *et al*. to image caption gastrointestinal tract examinations [10]. More recently transformer architectures have been proposed. Xiong *et al*. used a language model built on a transformer to generate captions for medical images while using a more naive image encoder architecture [34]. Selivanov *et al*. used transformers for generative medical image captioning [27], specifically using GPT-3 [4], and Show-Attend-Tell [35] as language models, and training on two medical image datasets Open-I [6], and MIMIC-CXR [13]. To the best of our knowledge, this is the first time that quantization techniques such as QLoRA have been attempted on the medical image captioning domain.

3. Methodology

To extend BLIP-2 to the medical image captioning domain, we first quantize BLIP-2 using QLoRA, and then

[1]For image encoders, they tried using ViT-L/14 from CLIP [24] and ViT-g/14 from EVA-CLIP [8]

[2]For LLMs, they tried using OPT [36] and FlanT5 [5]

fine-tune the quantized BLIP-2 on image-caption pairs in the ROCO dataset.

3.1. Medical Image-captioning Data

Since BLIP-2 was trained on data of common objects in natural settings, we use the medical image dataset Radiology Objects in COntext (ROCO) [23] to fine-tune our model. ROCO contains around 65K training, 8K validation, and 8K test radiology images, collected from PubMedCentral and ImageCLEF. The types of radiology images contained in the dataset include Computer Tomography (CT) scans, Magnetic Resonance Imaging (MRI), X-rays, and other medical imaging modalities. Each image is paired with its corresponding caption, which was created by a human annotator and extracted from a peer-reviewed biomedical research paper.

ROCO images and captions were pre-processed by the dataset creators to standardize the data. Compound figures from the original papers, where one figure contains several subfigures, were largely removed from the dataset. All images were resized to 360x360 pixels, using whitespace as a background. ROCO's ground-truth captions also omit stopwords and special characters.

3.2. Data Preprocessing

To preprocess the medical imaging data, we first downloaded the ROCO dataset, which is hosted by ROCO's authors on GitHub [23] [3].

We then create a HuggingFace Datasets object, containing a train-val-test split of images and their associated ground-truth captions. To tokenize both the image data and text data, we use the default tokenizer used by BLIP-2. The tokenizer also returns an attention mask associated with the text data prevent attention from being performed on padding token indices.

3.3. Pre-trained BLIP-2 model

We use BLIP-2's pre-trained model [17], which uses OPT-2.7b, a large language model that has 2.7 billion parameters. BLIP-2's model contains 3 components: (1) a frozen image-encoder, (2) a Querying Transformer (Q-FORmer), and a frozen LLM. Without quantization, BLIP-2's image captioning model has 1.1 billion trainable parameters. For image captioning, we specifically use the `Blip2ForConditionalGeneration` model.

3.4. Quantization

To quantize BLIP-2, we use HuggingFace's BitsAndBytes package. This package was developed by Dettmers *et al*. and was released with QLoRA. We follow the double quantization technique, and use 4-bit Normal Floats (NF4) as proposed in QLoRA. By applying quantization to each linear layer, we reduce the number of trainable parameters from around 3.7 billion to only 2.6 million, less than .07% of the original amount.

We only apply the QLoRA update matrices to the query and key projection layers of the language model's decoder self-attention module, which is noted as standard practice from HuggingFace.

3.5. Fine-tuning

To fine-tune BLIP-2 to extend BLIP-2 to the medical image caption domain, we fine-tune BLIP-2 on the training split of 65K images from the radiology data available in ROCO. We use the Parameter-Efficient Fine-Tuning (PEFT) framework available with QLoRA using HuggingFace's PEFT package, making sure to use 4-bit Normal Floats (NF4) and double quantization as proposed by QLoRA.[4] We keep the compute type as `torch.bfloat16` as suggested by QLoRA.

When fine-tuning, we use the optimizer AdamW and the following parameters:

- Epochs: 3
- Batch size: 8
- Learning rate: 1e-4

We trained with a batch size of 8 because this was the most images that we could fit on the GPU without running out of memory. We fine-tuned with 3 epochs because training with too many can cause overfitting, and too little we were afraid that the model would be able to learn. We used the common learning rate of 1e-4 as we did not have time to hyper-parameter tune.

We trained for about 17 hours on the Linux compute cluster Hyak using one 24GB GPU, provided for all University of Washington students.

3.6. Inference and Evaluation

When using the fine-tuned model to generate captions on the validation and test sets, we set the max length of generated captions to 50. Special tokens were skipped while decoding generated captions.

To evaluate the performance of our model, we use 4 different scores commonly used for image captioning: (1) Consensus-based Image Description Evaluation (CIDEr) [32], (2) Metric for Evaluation of Translation with Explicit ORdering (METEOR) [2], (3) Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [19], and (4) bilingual evaluation understudy (BLEU) [22], all computed using the

[3] Instructions to download ROCO can be found on GitHub (as of January 30, 2024).

[4] The PEFT package can be found on HuggingFace (as of January 30, 2024).

python package `pycocoevalcap`.[5] All four scores rely on n-gram matching between a target caption, usually written by a human annotator, and the generated caption.

4. Experiments & Results

4.1. Empirical Results

We compared our Quantized + Fine-tuned (QLoRA) BLIP-2 model with 2 different baselines, namely (Baseline1) a Not-Quantized + Pre-trained BLIP-2 model, as well as (Baseline2) a Quantized + Pre-trained BLIP-2 model. For all models, we ran inference on the validation data and test data provided by ROCO and reported all four metrics described in sec. (3.6).

For (Baseline1) Not-Quantized + Pre-trained BLIP-2 model we loaded in the pre-trained BLIP-2 model from HuggingFace without any quantization, and ran zero-shot inference out-of-the-box and did not perform any further fine-tuning.

For (Baseline2) Quantized + Pre-trained BLIP-2 model, we loaded in the quantized pre-trained BLIP-2 model from HuggingFace. To quantize the model, we loaded the model with 4-bit Normal Floats (NF4), using both NF4 and double quantization as proposed in QLoRA. We then ran zero-shot inference without any further fine-tuning.

From table 1, we see that our Quantized + Fine-tuned (QLoRA) model performed drastically better than the two aforementioned baselines. Our model more than doubled the CIDEr score of both baselines, as well as performed considerably better when compared then both baselines in regards to the three other evaluation metrics, METEOR, ROUGE, and BLEU.

4.2. Non-Empirical Results

The generated captions by the QLoRA BLIP-2 model consistently produce captions which feel more relevant than the two baseline models. The captions generated by the fine-tuned QLoRA BLIP-2 model consistently uses medical terms which both the baselines do not know. For instance, in image 4, we see that while quantized + pre-trained BLIP-2 (baseline2) mentions a "screw and a screwdriver", and non-quantized + pre-trained BLIP-2 also mentions a "screw", the caption generated by our QLoRA BLIP-2 model mentions an "postoperative", "implant in place", jargon which Baseline1 and Baseline2 are incapable of producing.

It is worth noticing that the baseline models also consistently identifies these images as "CT scans", "x-rays" or "an mri", while our fine-tuned QLoRA BLIP-2 model is capable of using using medical imaging vocabulary such as "angiogram" as seen in figure 2. In fact, the baselines are incapable of producing such a word and never once produce the word "aniogram". It is also relevant to note that our fine-tuned QLoRA BLIP-2 model also correctly identifies figure 2 as an aniogram, suggesting that our model has been able to distinguish aniograms from other types of medical imagery.

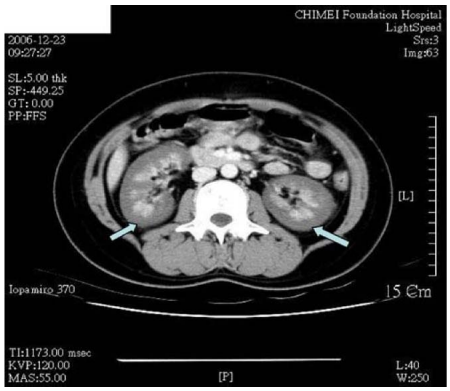

Figure 1. **Target caption**: "Contrast-enhanced computed tomography of the abdomen revealed absent opacification of the renal cortex and enhancement of subcapsular and juxtamedullary areas and the medulla without excretion of contrast medium"
Baseline1 generated caption: "a ct scan shows the location of the kidney and the kidney is shown in blue,"
Baseline2 generated caption: "a ct scan of the abdomen showing the location of the kidney"
QLoRA (ours) generated caption: "CT scan of the abdomen showing a large, heterogeneous, and heterogeneous mass in the right kidney."

5. Discussion & Future Work

5.1. Limitations

It is important to take the caption generation evaluation metrics with a grain of salt. Firstly, what constitutes a good image caption is subjective, and human annotators often will rank captions differently. Therefore, the target caption that the model is being evaluated against might not be the best representation of a caption for the given image, even while the target caption is treated as the best. Secondly, all four caption evaluation metric use some form of n-gram matching between the target and generated captions. This means that in order for a generated caption to receive a perfect score, the generated caption must match the target caption exactly. Therefore the score that an image-caption obtains is highly dependent on the presence or non-presence of tokens in the target captions, and that generated captions which are more verbose than the target captions are usually penalized, even if the verbosity of the generated captions provides more and better relevant information. With that being said, the four metrics generally correlate to the human understanding of a good caption, with higher metric scores correlating to a better reception by human annotators.

[5]The package can be found on GitHub (as of January 30, 2024).

Model	CIDEr		METEOR		ROUGE		BLEU	
	Val	Test	Val	Test	Val	Test	Val	Test
Not-Quantized + Pre-trained BLIP-2 (Baseline1)	0.328	0.352	0.073	0.073	0.123	0.125	0.040	0.040
Quantized + Pre-trained BLIP-2 (Baseline2)	0.036	0.384	0.074	0.075	0.129	0.130	0.041	0.041
Quantized + Finetuned (QLORA) BLIP-2 **(Ours)**	**0.786**	**0.765**	**0.084**	**0.084**	**0.178**	**0.178**	**0.092**	**0.091**

Table 1. The validation and

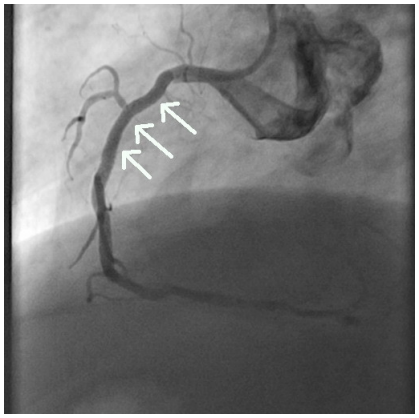

Figure 2. **Target caption**: "Coronary angiography image obtained just after stent placement in the right coronary artery showing the good revascularization of the artery."
Baseline1 generated caption: "a ct scan shows the heart and the arteries"
Baseline2 generated caption: "a ct scan of a patient with a large vein in the neck"
QLoRA (ours) generated caption: "Angiogram showing the left anterior descending artery (LAD) with a large stenosis (arrows) and the left anterior descending artery (LAD) with a small stenosis (arrowhead)."

Figure 3. **Target caption**: "Computed tomography scan of the chest showed presence of significant fat in the mediastinum surrounding the heart and other mediastinal structures."
Baseline1 generated caption: "a ct scan of a human lung with a small area of white"
Baseline2 generated caption: "a ct scan of a lung with a large tumor"
QLoRA (ours) generated caption: "CT scan of the chest showing a large left pulmonary nodule."

Not only are image captions hard for human annotators, but medical image captions are hard for human evaluators. Medical image captions contain jargon and vocabulary which is inaccessible to the general public, and take years of education to obtain. For this reason, the lack of medical experts who are able to provide quality feedback on the medical image captions generated by an image-captioner creates a bottleneck for the evaluation of the model. As non-medical professionals, it was incredibly hard for us to tell whether an image caption that we generated contained factual and pertinent information, which could make the generated captions better than they seem. For this reason, we highly relied on the four flawed image caption scores CIDEr, METEOR, ROUGE, and BLEU to evaluate our model.

Our model's performance is also constrained by the size of our dataset. While BLIP-2 was trained on 129M total images of objects and people in natural settings, our model was only further fine-tuned on 65K radiology images from ROCO. Even after finetuning, the model still has limited capabilities for captioning medical images.

5.2. Future Work

Our model could be extended for tasks like prompted image captioning, visual question answering, or chat-based prompting to further leverage its capabilities. As described in MedICaT, a paper presenting a separate medical image captioning dataset [31], text-image matching may be a useful extension of this model, where captions produced from the model are used to try to match user given captions to the original image. This could improve search systems within biomedical research paper databases, making the literature

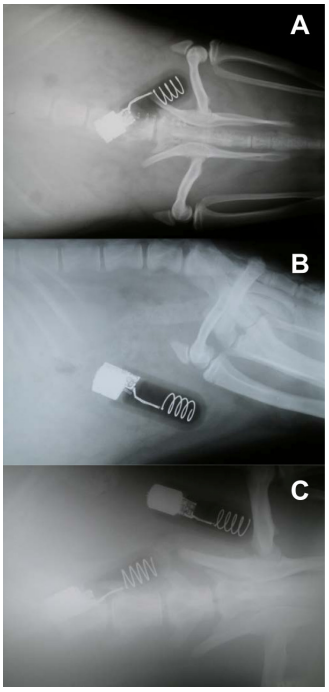

Figure 4. **Target caption**: "Post-surgery radiographs of California sea lions with implanted life history transmitters. (A) Dorsoventral view of single transmitter in animal CSL6018, a 66 kg female – the tag has a size of approximately three vertebrae. (B) Lateral view of single transmitter in CSL6018. (C) Dorsoventral view of dual transmitters in animal CSL6053, a 195 kg male. The tags have a size of approximately two vertebrae."
Baseline1 generated caption: "a picture of a broken arm with a screw in it,"
Baseline2 generated caption: "a x-ray of a hand with a screw and a screwdriver"
QLoRA (ours) generated caption: "Postoperative X-ray of the right wrist showing the implant in place."

review process more efficient for researchers.

Other future work would involve the use of medical professionals for the evaluation as well as creation of image captions. As medical jargon is inaccessible to the general public, medical professionals are necessary to ensure that the quality of image captions.

5.3. Ethical Considerations

It is important to be careful about the settings in which the model is used, especially in the medical field. The model should not be used in place of expert knowledge from trained medical professions for medical diagnosis. LLMs often hallucinate, may confidently produce verbose captions, though inaccurate outputs, which could have highly damaging implications if used directly for diagnosis on patients.

Another important consideration is the method of data collection. Because the dataset is made up of medical images and corresponding captions, it is important not to include any personally identifying information (PID) in both the image and captions. It also is important that the patients have given informed consent before distributing their medical images to others.

6. Conclusion

In conclusion, we extend the use of the quantization technique QLoRA of PEFT models to that of the medical image captioning domain. We saw that the use of QLoRA in conjunction with BLIP-2 yielded results which makes considerable improvements over the baselines, and that the usage of quantized PEFT models provides us with ways that we can scale state-of-the-art pre-trained image captioning models to different image captioning domains using very little compute.

Although QLoRA was originally applied to fine-tuning on LLMs, our application of QLoRA and quantization methods gives one example of how QLoRA can be applied to other domains like medical image captioning. Our results show that tools like image captioning can be made accessible for other research areas, such as medical research, even in low resource settings with memory limitations.

References

[1] Viktar Atliha and Dmitrij Šešok. Image-captioning model compression. *Applied Sciences*, 12(3), 2022. 2

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. 3

[3] Djamila-Romaissa Beddiar, Mourad Oussalah, and Tapio Seppänen. Automatic captioning for medical imaging (mic): a rapid review of literature. *Artificial Intelligence Review*, 56(5):4019–4076, May 2023. 2

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 2

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun

Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. 2

[6] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 07 2015. 2

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. 1, 2

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 2

[9] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning, 2017. 2

[10] Philipp Harzig, Moritz Einfalt, and Rainer Lienhart. Automatic disease detection and report generation for gastrointestinal tract examination. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2573–2577, New York, NY, USA, 2019. Association for Computing Machinery. 2

[11] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning, 2018. 2

[12] Weitao Jiang, Xiying Li, Haifeng Hu, Qiang Lu, and Bohong Liu. Multi-gate attention network for image captioning. *IEEE Access*, 9:69700–69709, 2021. 2

[13] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. 2

[14] Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization, 2023. 2

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016. 1

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. 2

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1, 2, 3

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2

[19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1

[21] Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models, 2023. 2

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. Association for Computational Linguistics. 3

[23] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 11043:180–189, 2018. 1, 3

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[25] Harshit Rampal and Aman Mohanty. Efficient cnn-lstm based image captioning using neural network compression, 2020. 2

[26] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 1

[27] Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V. Dylov. Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1):4171, Mar 2023. 2

[28] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. Image captioning: A comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pages 325–328, 2020. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2

[30] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning, 2021. 2

[31] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer

Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references, 2020. 5

[32] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 3

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015. 2

[34] Yuxuan Xiong, Bo Du, and Pingkun Yan. Reinforced transformer for medical image captioning. In *MLMI@MICCAI*, 2019. 2

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016. 2

[36] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. 2

[37] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models, 2023. 2