OH! WE FREEZE: IMPROVING QUANTIZED KNOWLEDGE DISTILLATION VIA SIGNAL PROPAGATION ANALYSIS FOR LARGE LANGUAGE MODELS

Kartikeya Bhardwaj*, **Nilesh Prasad Pandey***, **Sweta Priyadarshi, Kyunggeun Lee, Jun Ma, Harris Teague**
Qualcomm AI Research†
{kbhardwa,nileshpr,swetpriy,kyunggeu,jma,hteague}@qti.qualcomm.com

ABSTRACT

Large generative models such as large language models (LLMs) and diffusion models have revolutionized the fields of NLP and computer vision respectively. However, their slow inference, high computation and memory requirement makes it challenging to deploy them on edge devices. In this study, we propose a light-weight quantization aware fine tuning technique using knowledge distillation (KD-QAT) to improve the performance of 4-bit weight quantized LLMs using commonly available datasets to realize a popular language use case, on device chat applications. To improve this paradigm of finetuning, as main contributions, we provide insights into stability of KD-QAT by empirically studying the gradient propagation during training to better understand the vulnerabilities of KD-QAT based approaches to low-bit quantization errors. Based on our insights, we propose *ov-freeze*, a simple technique to stabilize the KD-QAT process. Finally, we experiment with the popular 7B LLaMAv2-Chat model at 4-bit quantization level and demonstrate that ov-freeze results in near floating point precision performance, i.e., less than 0.7% loss of accuracy on Commonsense Reasoning benchmarks.

1 INTRODUCTION

The increasing popularity of large generative neural networks, LLaMA (Touvron et al. (2023), OPT Zhang et al. (2022)) and diffusion models (Dhariwal & Nichol (2021); Ho et al. (2020)), has revolutionized the field of machine learning exhibiting exceptional capabilities in generating realistic human-like text and images. However, due to large compute and memory requirements of these models, deploying these models on resource-constrained devices is challenging. Among various compression methods available, quantization (Nagel et al. (2021); Krishnamoorthi (2018); Jacob et al. (2018)) has proven to be a promising technique to optimize various vision and language models for deployment on resource-constrained devices.

In this paper, we focus on INT4 uniform, static, channelwise weight and tensorwise activation quantization of LLMs. Most quantization literature uses dynamic or blockwise methods for 4-bit or lower quantization to recover accuracy. However, such quantization schemes may not be supported on Neural Processing Units (NPU) hardware. Channelwise weight and tensorwise activation quantization is significantly more easy to support in terms of hardware and software implementation in commercial NPUs. Keeping such strong constraints in mind, when we perform INT4 quantization of weights, quantization noise increases significantly which risks destabilizing the entire training dynamics. Therefore, we need to analyze vulnerabilities in forward and backward passes within quantized LLMs. Hence, improving accuracy for uniform, static, channelwise weight and tensorwise activation quantization is extremely important for deployment use cases of LLMs.

To this end, we propose to *empirically* analyze the signal propagation through a well-known chat use case LLM, LLaMAv2-Chat, to understand its vulnerabilities to quantization errors. Specifically, we focus on multi-head self-attention modules of LLaMAv2-Chat (see Fig. 1) and conduct a detailed analysis of its forward and backward pass signals to see which components of the attention modules can cause KD-QAT to destabilize. After identifying these vulnerabilities, we propose specific

*Equal Contribution. †Qualcomm AI Research is an initiative of Qualcomm Technologies Inc.

오! 우리는 얼어붙는다: 신호 전파 분석을 통한 대형 언어 모델의 양화 지식 증류 개선

카르티케야 바르다와지*, 니레시 프라사드 판데이*, 스베타 프리야다르시, 경근 리, 준 마, 해리스 티그
퀄컴 AI 연구†{kbhardwa,nileshpr,swetpriy,kyunggeu,jma,hteague}@qti.qualcomm.com

요약

대규모 생성 모델인 대규모 언어 모델(LLM)과 확산 모델은 각각 NLP와 컴퓨터 비전 분야를 혁신했습니다. 그러나 느린 추론 속도, 높은 연산 및 메모리 요구 사항으로 인해 엣지 장치에 배포하는 것이 어렵습니다. 본 연구에서는 지식 증류(KD-QAT)를 활용한 경량 양자화 인식 미세 조정 기술을 제안하여 일반적으로 사용되는 데이터셋을 사용하여 4비트 가중치 양자화된 LLM의 성능을 개선하고, 장치 내 채팅 애플리케이션과 같은 인기 있는 언어 사용 사례를 구현합니다. 이 미세 조정 패러다임을 개선하기 위해, 주요 기여로 KD-QAT의 안정성에 대한 통찰력을 제공하고, 훈련 중 기울기 전파를 경험적으로 연구하여 KD-QAT 기반 접근 방식이 저비트 양자화 오류에 취약한 점을 더 잘 이해합니다. 이러한 통찰력에 따라, KD-QAT 과정을 안정화하기 위한 간단한 기술인 *ov-freeze*를 제안합니다. 마지막으로, 인기 있는 7B LLaMAv2-Chat 모델을 4비트 양자화 수준에서 실험하고, Commonsense Reasoning 벤치마크에서 부동 소수점 정밀도 성능에 근접하는 결과를 얻었습니다. 즉, 정확도 손실이 0.7% 미만임을 보였습니다.

1 소개

대형 생성적 신경망의 인기가 높아짐에 따라 LLaMA (Touvron et al. (2023), OPT Zhang et al. (2022))와 확산 모델 (Dhariwal & Nichol (2021); Ho et al. (2020))이 기계 학습 분야를 혁신하고 있으며, 현실적이고 인간과 유사한 텍스트와 이미지를 생성하는 뛰어난 능력을 보여주고 있습니다. 그러나 이러한 모델은 많은 계산과 메모리 자원을 필요로 하기 때문에 자원 제약이 있는 장치에 배포하는 것이 어렵습니다. 다양한 압축 방법 중에서도 양화(Nagel et al. (2021); Krishnamoorthi (2018); Jacob et al. (2018))는 비전 및 언어 모델을 최적화하여 자원 제약이 있는 장치에 배포하는 데 효과적인 기술로 입증되었습니다.

이 논문에서는 LLM의 INT4 균일, 정적, 채널별 가중치 및 텐서별 활성화 양자에 초점을 맞춥니다. 대부분의 양자화 문헌은 4비트 이하 양자화를 복구하기 위해 동적 또는 블록별 방법을 사용합니다. 그러나 이러한 양자화 방식은 신경 처리 장치(NPU) 하드웨어에서 지원되지 않을 수 있습니다. 채널별 가중치와 텐서별 활성화 양자는 상업용 NPU에서 하드웨어 및 소프트웨어 구현 측면에서 훨씬 더 쉽게 지원됩니다. 이러한 강력한 제약 조건을 염두에 두고 가중치에 대한 INT4 양자화를 수행할 때 양자 잡음이 크게 증가하여 전체 학습 동역학을 불안정하게 만들 위험이 있습니다. 따라서 양자화된 LLM 내에서 전방 및 후방 패스의 취약점을 분석해야 합니다. 따라서 균일하고 정적이며 채널별 가중치와 텐서별 활성화 양자의 정확도를 향상시키는 것은 LLM의 배포 사용 사례에 매우 중요합니다.

이 목적을 위해, 우리는 잘 알려진 채팅 사용 사례 LLM인 LLaMAv2-Chat를 통해 신호 전파를 *empirically* 분석하여 양화 오류에 대한 취약점을 이해할 것을 제안합니다. 구체적으로, 우리는 LLaMAv2-Chat의 다중 헤드 자기 주의 모듈(그림 1 참조)에 초점을 맞추고, 주의 모듈의 어떤 구성 요소가 KD-QAT를 불안정하게 만들 수 있는지 확인하기 위해 순전파 및 역전파 신호에 대한 상세한 분석을 수행합니다. 이러한 취약점을 식별한 후, 우리는 구체적인 해결책을 제안합니다.

*동등한 기여 †Qualcomm AI Research is an initiative of Qualcomm Technolo지스 주식회사 1

solutions to stabilize the KD-QAT process to achieve significant improvements in INT4 quantized LLaMAv2-Chat network accuracy enabling its deployment on commercially available hardware.

We make the following **key contributions** in this paper:

1. We create a simple pipeline for KD-QAT for a well-known chat use case LLM, LLaMAv2-Chat-7B model. It uses public datasets and requires less than a day to provide INT4 quantized networks on a single node containing 8 NVIDIA A100 GPUs.
2. We analyze forward and backward pass in detail for this network and find that low-bit quantization non-uniformly impacts different parts of the attention modules. Specifically, o- and v-projection layers are more susceptible to quantization errors than other weight layers in the multi-head self-attention module.
3. Based on our analysis, we propose *ov-freeze* (oh! we freeze). It shows the importance of properly analyzing layers susceptible to low-bit quantization before embarking on quantization-aware-training. We demonstrate that this improves accuracy significantly across multiple benchmarks and closes the gap between quantized and FP models.

2 Related Work

Many methods have been proposed in the model efficiency space to solve the problem of neural network quantization. Quantization not only reduces model size but also leverages efficient fixed-point representation over floating-point representations. Most quantization techniques can be categorized either as post-training quantization (PTQ) (Nagel et al. (2020; 2019); Dong et al. (2019)) or quantization-aware-training (QAT)(Bhalgat et al. (2020); Esser et al. (2019); Nagel et al. (2022)) methods methods. Although, PTQ are go-to compression techniques (Frantar et al. (2023); Yao et al. (2022)) for most generative models, these methods suffer huge degradation in performance post quantization for low bitwidth quantization. QAT based approaches are difficult for these models due to lack of access to high quality data and training recipes required for finetuning these models.

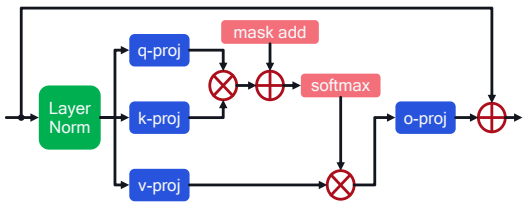

Figure 1: Multi-Head Self-Attention modules in LLaMAv2-Chat model. We focus on the forward and backward signal propagation properties after the q-, k-, v-, and o-projection layers to understand their vulnerabilities to low-bit quantization.

In this work, we propose to use knowledge distillation to finetune weight only quantized language models to achieve near floating point performance. LLM-QAT (Liu et al. (2023)) takes a similar approach but uses teacher generated dataset for finetuning, which is not always feasible due to sensitivity of generated dataset on sampling strategies and compute and memory intensive nature of the generation process. In this work, we use publicly available datasets for the distillation process, and perform ablations to understand vulnerabilities in the multi-head self-attention modules of the LLaMAv2-Chat model.

3 Method

In this section, we first describe our quantization and KD-QAT setup and analyze the gradient values of the LLaMAv2-Chat model to understand low-bit quantization vulnerabilities in the multi-head self-attention modules. Then, based on this analysis, we propose ov-freeze, a technique to stabilize our KD-QAT training pipeline.

3.1 Preliminaries

While quantizing neural networks, real valued weight W^r and activations x^r are quantized to low precision value. Hence, for a b bitwidth uniform quantizer with scale s and zero-point z, asymmetric quantization is defined as

$$q(W^r; s, z) = s \cdot \left[clamp(\left\lfloor \frac{W^r}{s} \right\rceil + z, 0, 2^b - 1) - z \right] \tag{1}$$

KD-QAT 프로세스를 안정화하여 INT4 양화 LLaMAv2-Chat 네트워크의 정확도를 크게 향상시키는 솔루션을 구현하여 상업적으로 이용 가능한 하드웨어에 배포할 수 있도록 합니다.

이 논문에서 우리는 다음과 같은 주요 기여를 합니다:

1. 잘 알려진 채팅 사용 사례 LLM인 LLaMAv2-Chat-7B 모델을 위한 간단한 KD-QAT 파이프라인을 생성합니다. 이 파이프라인은 공개 데이터셋을 사용하고 8개의 NVIDIA A100 GPU가 있는 단일 노드에서 하루 미만의 시간 내에 INT4 정량화된 네트워크를 제공합니다. 2. 이 네트워크의 순방향 및 역방향 전달을 자세히 분석하여 저비트 정량화가 주의력 모듈의 서로 다른 부분에 비균일하게 영향을 미친다는 것을 발견합니다. 특히, o- 및 v-투영 층은 멀티 헤드 자기 주의 모듈의 다른 가중치 층보다 정량화 오류에 더 취약합니다. 3. 분석을 바탕으로 *ov-freeze* (우리는)를 동결을 제안합니다. 이는 저비트 정량화에 취약한 층을 제대로 분석하는 것이 정량화 인식 훈련을 시작하기 전에 중요하다는 것을 보여줍니다. 이 방법이 여러 벤치마크에서 정확도를 크게 향상시키고 정량화 및 FP 모델 간의 격차를 해소한다는 것을 보여줍니다.

2 관련 연구

신경망 양화에 대한 문제를 해결하기 위해 모델 효율성 영역에서 많은 방법들이 제안되었습니다. 양화는 모델 크기를 줄일 뿐만 아니라 부동 소수점 표현에 비해 효율적인 고정 소수점 표현을 활용합니다. 대부분의 양화 기술은 사후 훈련 양화(PTQ) (Nagel 외, 2020, 2019; Dong 외, 2019) 또는 양화 인식 훈련(QAT) (Bhalgat 외, 2020; Esser 외, 2019; Nagel 외, 2022) 방법 중 하나로 분류할 수 있습니다. 비록 PTQ가 대부분의 생성 모델에 대한 일반적인 압축 기술(Frantar 외, 2023; Yao 외, 2022)이지만, 이러한 방법들은 저 비트 너비 양화 후 성능이 크게 저하됩니다.

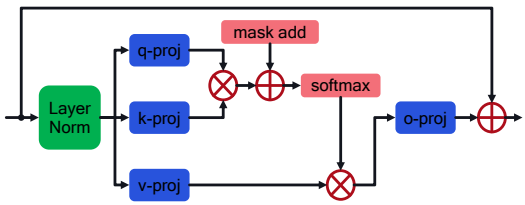

그림 1: LLaMAv2-Chat 모델의 Multi-Head Self-Attention 모듈. 저희는 q-, k-, v-, o-투영 층 이후의 전방 및 후방 신호 전파 특성에 초점을 맞추어 저비트 양자에 대한 취약성을 이해합니다.

정량화. QAT 기반 접근 방식은 이러한 모델에 대해 고품질 데이터 및 미세 조정하기 위한 훈련 레시피에 대한 접근성이 부족하여 어려움을 겪습니다.

이 연구에서는 지식 증류를 사용하여 가중치만 양화된 언어 모델을 미세 조정하여 부동 소수점 성능에 가까운 성능을 달성하는 것을 제안합니다. LLM-QAT(Liu et al., 2023)는 유사한 접근 방식을 취하지만, 미세 조정을 위해 교사 생성 데이터셋을 사용하는데, 이는 생성 데이터셋의 샘플링 전략에 대한 민감성과 생성 과정의 계산 및 메모리 집약적인 특성 때문에 항상 가능한 것은 아닙니다. 이 연구에서는 공개적으로 접근 가능한 데이터셋을 증류 과정에 사용하고, LLaMAv2-Chat 모델의 다중 헤드 자기 주의 모듈의 취약점을 이해하기 위해 제거 실험을 수행합니다.

3 방법

이 섹션에서는 먼저 양자화 및 KD-QAT 설정을 설명하고 LLaMAv2-Chat 모델의 기울기 값을 분석하여 다중 헤드 자기 주의 모듈의 저비트 양자화 취약점을 이해합니다. 그런 다음 이 분석을 바탕으로 KD-QAT 훈련 파이프라인을 안정화하기 위한 ov-freeze라는 기술을 제안합니다.

3.1 예비사항

신경망을 양자화할 때, 실수 값의 가중치 W^r와 활성화 x^r는 저정밀도 값으로 양자화됩니다. 따라서 b 비트 너비 균일 양자화기(scale s 및 제로 포인트 z)에 대해 비대칭 양자화는 다음처럼 정의됩니다.

$$q(W^r; s, z) = s \cdot \left[clamp(\left\lfloor \frac{W^r}{s} \right\rceil + z, 0, 2^b - 1) - z \right] \quad (1)$$

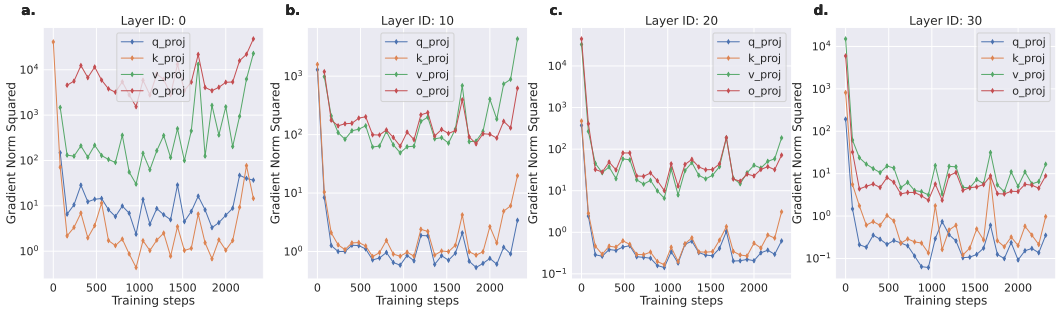

Figure 2: Relative comparison of gradient norm squared values between q-, k-, v-, o-projection outputs at various LLaMA self-attention modules during KD-QAT: (a) Layer 0, (b) Layer 10, (c) Layer 20, and (d) Layer 30. Clearly, o- and v- projection gradient norm squared values are one or two orders of magnitude higher than that of q- and k- projection layers. Such higher gradient values may destabilize o- and v-layers faster than q- and k-projections during KD-QAT, especially for low-bit quantization (e.g., INT4).

For our experiments, we use AIMET [1] (Siddegowda et al. (2022)) to quantize the models to desired bit-widths and use per-channel symmetric quantization for weights and asymmetric quantization for activations respectively. We set the quantization range of each weight quantizer using a mean squared error (MSE) based criteria minimizing the local MSE between FP and quantized tensors.

3.2 KD-QAT Setup

To demonstrate the effectiveness of our KD-QAT methodology on an established chat application, we experiment with the popular LLaMAv2-Chat model (Touvron et al. (2023)), and quantize it to W4A16[2]. For knowledge distillation based QAT, we use full-precision chat model as the teacher and quantized model as our student. Contrary to LLM-QAT (Liu et al. (2023)), we observe that using a weighted sum of cross-entropy and KL divergence for training leads to more stable and better convergence.

3.3 Gradient Properties of Quantized LLM

We investigate the gradient values of various components of the LLaMAv2-Chat multi-head self-attention modules (see Fig. 1) to understand which layers may contribute to destablization when we conduct INT4 weight quantization. For layer activations with output feature map, Θ, and task loss $\mathcal{L}$, we compute the frobenius norm squared of the gradients at this layer's output:

$$||g||_2^2 = ||\nabla_\Theta \mathcal{L}(\Theta)||_2^2 \tag{2}$$

In Fig. 2, we show the comparisons among these gradient norm squared values for q-, k-, v-, and o-projection outputs at different hidden modules[3]. As evident, the gradient norm squared for o- and v-projections are one or two orders of magnitude higher than that for q- and k-layers. Moreover, as shown in Fig. 2 (a, b), the v-layers (green lines) in early hidden modules (Layer ID: 0 and 10) seem to have largest variation in gradient norm compared to other layers. Therefore, when we perform low-bit (e.g., INT4) quantization, these observations suggest that o- and v-layers are much more vulnerable to quantization errors than q- and k- layers. Specifically, results in Fig. 2 indicate that significantly higher gradient values may create bigger, more abrupt changes in o- and v- layers which would more directly impact their forward pass outputs. Therefore, forward pass of o- and v-projections may exhibit unstable behavior due to such high gradient values.

3.4 Oh! We Freeze

Since o- and v-projection layers can be senstive to sudden parameter changes due to high gradients, these layers can destabilize the entire KD-QAT training trajectory. Therefore, we propose *ov-freeze*

[1] AIMET is a product of Qualcomm Innovation Center, Inc., available on GitHub at `https://github.com/quic/aimet`

[2] WxAy quantization indicates quantizing all the weights and output activations to x and y bits respectively.

[3] Hidden modules are labeled as Layer IDs throughout this paper. For instance, "Layer ID: 10" refers to the 10th hidden module out of the total 32 modules containing multi-head self-attentions in LLaMAv2 model.

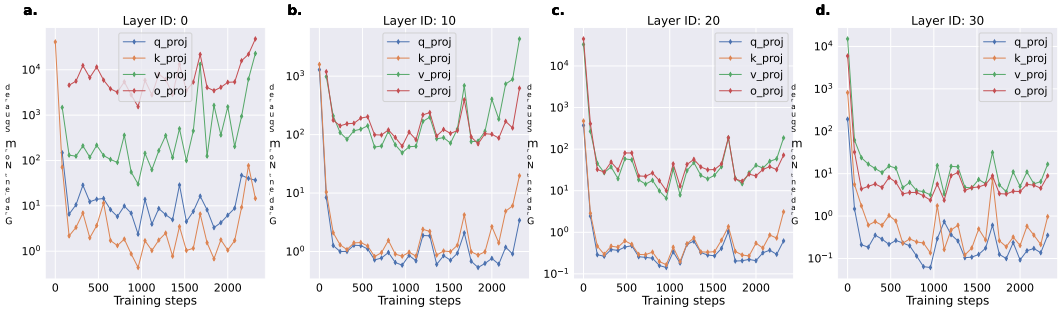

그림 2: KD-QAT 중 다양한 LLaMA 자기 주의 모듈의 q-, k-, v-, o-투영 출력 간의 기울기 노름 제곱 값의 상대 비교: (a) 레이어 0, (b) 레이어 10, (c) 레이어 20 및 (d) 레이어 30. 명확히 o- 및 v-투영 기울기 노름 제곱 값은 q- 및 k-투영 계층보다 1~2단계 높은 순서이다. 이러한 높은 기울기 값은 특히 INT4와 같은 저비트 양자로화에서 KD-QAT 중 o- 및 v-계층이 q- 및 k-투영보다 더 빨리 불안정해질 수 있다.

저희 실험을 위해, 저희는 AIMET [1] (Siddegowda 외 (2022))를 사용하여 모델을 원하는 비트 너비로 양자화하고, 가중치에는 채널별 대칭 양자화를, 활성화에는 비대칭 양자화를 각각 사용합니다. 각 가중치 양자화의 양자화 범위는 FP(부동 소수점) 텐서와 양자화된 텐서 사이의 지역 MSE(평균 제곱 오차)를 최소화하는 MSE 기반 기준을 사용하여 설정합니다.

3.2 KD-QAT 설정

우리의 KD-QAT 방법론의 효과를 입증하기 위해 확립된 채팅 애플리케이션에서 인기 있는 LLaMAv2-Chat 모델(Touvron et al. (2023))을 실험하고 이를 W4A16[2]로 양자화합니다. 지식 증류 기반 QAT를 위해 전체 정밀도 채팅 모델을 교사(teacher)로, 양자화된 모델을 학생(student)으로 사용합니다. LLM-QAT(Liu et al. (2023))와 달리, 교차 엔트로피와 KL 발산(divergence)의 가중 합을 사용하여 훈련하는 것이 더 안정적이고 나은 수렴으로 이어진다는 것을 관찰합니다.

3.3 양자화된 LLM의 그래디언트 특성

우리는 LLaMAv2-Chat 다중 머리 자기 주의 모듈(그림 1 참조)의 다양한 구성 요소의 기울기 값을 조사하여 INT4 가중치 양자화를 수행할 때 불안정화에 기여할 수 있는 레이어를 이해합니다. 출력 특징 지도 Θ와 작업 손실 $\mathcal{L}$가 있는 레이어 활성화에 대해 이 레이어 출력에서의 기울기의 프로벤리우스 노름 제곱을 계산합니다:

$$||g||_2^2 = ||\nabla_\Theta \mathcal{L}(\Theta)||_2^2 \tag{2}$$

그림 2에서, 우리는 q-, k-, v-, 및 o-투영 출력에서의 기울기 노름 제곱 값의 비교를 다른 숨겨진 모듈[3]에서 보여준다. 명확하게, o- 및 v-투영에 대한 기울기 노름 제곱은 q- 및 k-층에 비해 1~2승의 크기가 더 크다. 또한, 그림 2(a, b)에서 보듯이, 초기 숨겨진 모듈(층 ID: 0 및 10)에서의 v-층(녹색 선)은 다른 층에 비해 기울기 노름에서 가장 큰 변이를 보이는 것 같다. 따라서, 저비트(예: INT4) 양자를 수행할 때, 이러한 관찰은 o- 및 v-층이 q- 및 k-층보다 양자화 오류에 훨씬 더 취약하다는 것을 시사한다. 구체적으로, 그림 2의 결과는 크게 높은 기울기 값이 o- 및 v-층에서 더 크고 갑작스러운 변화를 일으킬 수 있음을 나타내며, 이는 그들의 순방향 전달 출력에 더 직접적인 영향을 미칠 것이다. 따라서, o- 및 v-투영의 순방향 전달은 이러한 높은 기울기 값으로 인해 불안정한 행동을 보일 수 있다.

3.4 오! 우리는 얼어붙는다

o- 및 v-투영 레이어는 높은 기울기 때문에 갑작스러운 매개변수 변화에 민감할 수 있으며, 이러한 레이어는 전체 KD-QAT 훈련 궤도를 불안정하게 만들 수 있습니다. 따라서 우리는 *ov-freeze*을 제안합니다.

[1] AIMET is a product of Qualcomm Innovation Center, Inc., available on GitHub at `https://github.com/quic/aimet`

[2] WxAy quantization indicates quantizing all the weights and output activations to x and y bits respectively.

[3] Hidden modules are labeled as Layer IDs throughout this paper. For instance, "Layer ID: 10" refers to the 10th hidden module out of the total 32 modules containing multi-head self-attentions in LLaMAv2 model.

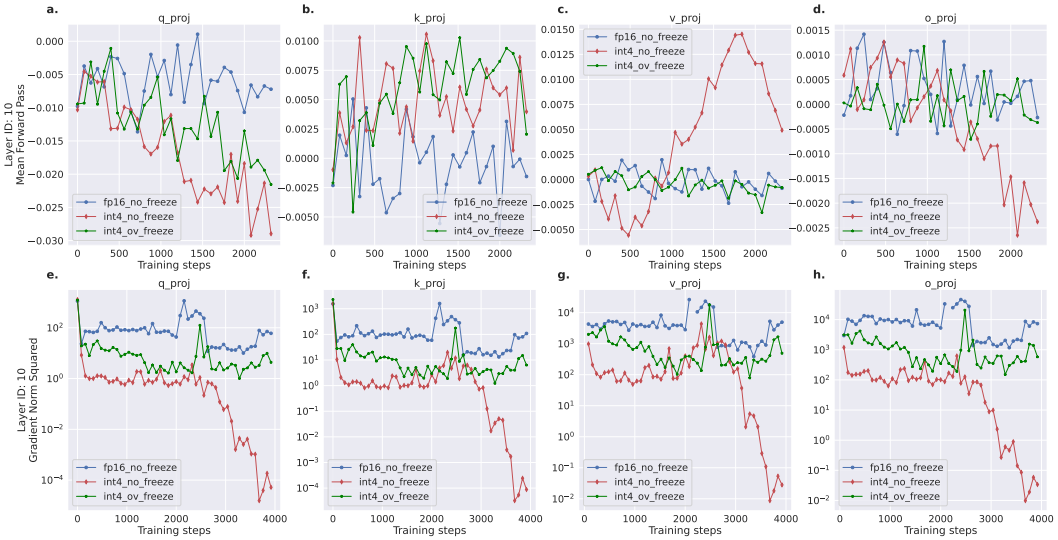

Figure 3: Forward and backward pass signal propagation analysis on a multi-head self-attention module at one of the hidden LLaMAV2-chat layers. We first analyze forward pass by tracking the mean value of the forward pass at the output of (a) q-projection, (b) k-projection, (c) v-projection, and (d) o-projection layers. Clearly, o and v layers are the most unstable as evident from the INT4 KD-QAT, no-freeze run. With ov-freeze, the forward pass behavior of the INT4 quantized network resembles that of the floating point model. Next, we analyze the backward pass behavior at the output of (e) q-projection, (f) k-projection, (g) v-projection, and (h) o-projection layers. Again, ov-freeze makes gradients more similar to the floating point training.

to stabilize the KD-QAT training process. As the name suggests, ov-freeze fixes the o- and v-projection weights to their post-training quantization values, freezes them and trains the rest of the network. This results in model adapting to quantization error without modifying the most sensitive parts of the network. Note that, v- and o-layers constitute a much smaller part of the complete LLM since majority of the parameters are contained within the MLP layers. In the next section, we will demonstrate how the proposed method significantly stabilizes the training loop and improves the final quantization results.

4 Experiments

In this section, we first describe the training setup and datasets. Then, we conduct extensive experiments to show that ov-freeze stabilizes both the KD-QAT training forward and backward passes. Finally, we present results demonstrating the superior W4A16 quantized accuracy obtained using ov-freeze on several benchmark tasks.

4.1 Training Setup and Datasets

To realize an important language use case, on device chat applications, we perform KD-QAT experiments on the popular 7B LLaMAv2-Chat. We used a 660B-token language dataset for the finetuning process. To evaluate the performance of the quantized networks, we use Wikitext and the CommonSense benchmarks - PIQA, Arc-Easy, Arc-Challenge, WinoGrande, OpenbookQA and Hellaswag.

4.2 Signal Propagation in LLaMAv2-Chat KD-QAT

4.2.1 Forward Pass Analysis

As the very first metric, we analyze the mean value of output activations of certain layers in the network. Specifically, Fig. 3 (a-d) show how the mean output activation value of q-, k-, v-, and o-projections vary across thousands of training steps for a specific LLaMAv2 hidden module. Here, we have used the FP16-no-freeze model (blue line) as a reference model. As evident from the INT4-no-freeze experiment (red line), o- and v-projection outputs deviate significantly from the FP16 reference. In contrast, when we freeze these two layers in all hidden modules (ov-freeze, green line), the mean forward pass follows the FP16 reference closely. Note that, even though we freeze only the o- and v-layers, the forward pass behavior still improves for other layers like the

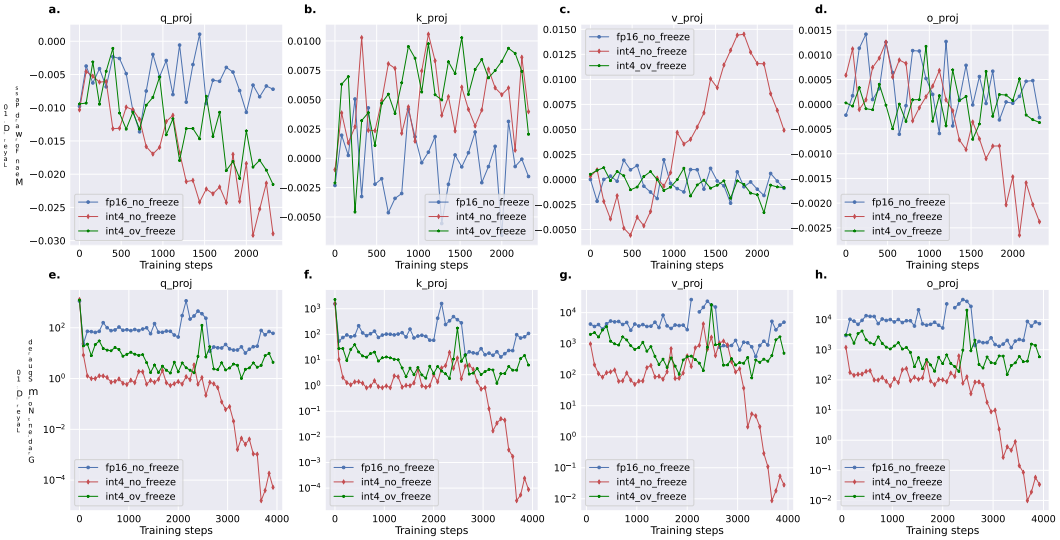

그림 3: LLaMAV2-chat 숨겨진 레이어 중 하나에서 다중 헤드 자기 주의 모듈의 전방 및 후방 패스 신호 전파 분석. 먼저 (a) q-투영, (b) k-투영, (c) v-투영 및 (d) o-투영 레이어의 출력에서 전방 패스의 평균 값을 추적하여 전방 패스를 분석합니다. 분명히, o 및 v 레이어는 INT4 KD-QAT, no-freeze 실행에서 나타난 것처럼 가장 불안정합니다. ov-freeze로 INT4 정량화된 네트워크의 전방 패스 동작은 부동 소수점 모델과 유사해집니다. 다음으로 (e) q-투영, (f) k-투영, (g) v-투영 및 (h) o-투영 레이어의 출력에서 후방 패스 동작을 분석합니다. 다시 한 번, ov-freeze는 기울기를 부동 소수점 학습과 더 유사하게 만듭니다.

KD-QAT 훈련 과정을 안정화하기 위해. 이름에서 알 수 있듯이 ov-freeze는 o- 및 v- 투영 가중치를 훈련 후 양자화 값에 고정시키고 이를 얼려서 네트워크의 나머지 부분만 훈련합니다. 이는 모델이 네트워크의 가장 민감한 부분을 수정하지 않고 양자화 오류에 적응하도록 합니다. MLP 계층에 대부분의 매개변수가 포함되어 있어 v- 및 o-계층이 전체 LLM에서 훨씬 작은 부분을 차지한다는 점에 유의하십시오. 다음 섹션에서는 제안된 방법이 어떻게 훈련 루프를 크게 안정시키고 최종 양자화 결과를 개선하는지 보여 드리겠습니다.

4 실험

이 섹션에서는 먼저 훈련 설정과 데이터셋을 설명합니다. 그런 다음 ov-freeze가 KD-QAT 훈련 전방 및 후방 전달을 모두 안정화한다는 것을 보여주기 위해 광범위한 실험을 수행합니다. 마지막으로 ov-freeze를 사용하여 여러 벤치마크 작업에서 우수한 W4A16 양자화 정확도를 얻은 결과를 제시합니다.

4.1 훈련 설정 및 데이터셋

중요한 언어 사용 사례를 실현하기 위해, 기기 내 채팅 애플리케이션에서 우리는 인기 있는 7B LLaMAv2-Chat에 대한 KD-QAT 실험을 수행합니다. 미세 조정 프로세스에는 660B 토큰 언어 데이터 세트를 사용했습니다. 정량화된 네트워크의 성능을 평가하기 위해 Wikitext 및 공통 감각 벤치마크 - PIQA, Arc-Easy, Arc-Challenge, WinoGrande, OpenbookQA 및 Hellaswag를 사용합니다.

4.2 시그널 전파 라마V2-챗 KD-QAT에서

4.2.1 전향적 분석

처음으로 분석하는 지표로서, 우리는 네트워크의 특정 레이어에서 출력 활성화의 평균 값을 분석합니다. 구체적으로, 그림 3 (a-d)는 특정 LLaMAv2 숨은 모듈에서 수천 번의 학습 단계 동안 q-, k-, v-, 그리고 o-투영에 대한 평균 출력 활성화 값이 어떻게 변하는지를 보여줍니다. 여기서 우리는 FP16-no-freeze 모델(파란색 선)을 기준 모델로 사용했습니다. INT4-no-freeze 실험(빨간색 선)에서 명확하게 알 수 있듯이, o- 및 v-투영 출력은 FP16 기준에서 크게 벗어납니다. 반면, 이 두 레이어를 모든 숨은 모듈에서 동결할 때(ov-freeze, 초록색 선), 평균 순전파 과정은 FP16 기준을 밀접하게 따릅니다. 주목할 점은, 오직 o- 및 v-레이어만 동결하더라도 다른 레이어들의 순전파 동작이 여전히 개선된다는 것입니다.

q-projection output shown in Fig. 3(a). These instabilities in forward pass can easily destabilize the training trajectory during KD-QAT. Note that, although Fig. 3 shows results for hidden module (Layer ID) 10, we observed similar trends for other hidden modules as well (see Appendix A).

Next, we empirically explore the impact of ov-freeze on backward pass.

4.2.2 Backward Pass Analysis

We now analyze the gradients at the outputs of q-, k-, v-, and o-projection layers at the same hidden module from the last section. We analyze gradient norm squared values as shown in Fig. 3 (e-h). Clearly, the INT4 ov-freeze (green line) again demonstrates similar gradient behavior as the reference FP16-no-freeze training run (blue line). In contrast, the INT4 no-freeze experiment (red line) has dramatically different backpropagation characteristics compared to the reference. All these instabilities result in lower final accuracy and an unstable KD-QAT.

4.3 Results

In this section, we compare performances of FP16 and quantized models obtained by our proposed freezing scheme on multiple benchmarks. First, Table 1 reports perplexity of PTQ and KD-QAT optimized 4-bit weight only quantized LLaMAv2-Chat on the Wikitext dataset. In our ablation, we consider different freezing schemes, o-, v-, ov-, qkv- and qkvo-, and observe freezing layers with relatively higher gradient norms values, i.e. o- and v-, along with other layers help significantly in improving both the stability of the KD-QAT training and the perplexity on Wikitext, achieving at-par or better performance than the FP model. As shown in Table 1, freezing ov- layers outperforms all other freezing schemes and PTQ baselines achieving better Wikitext perplexity than the FP16 model.

Model	Perplexity (ppl↓)
FP16	7.08
PTQ: Min-Max range setting	9.69 (+2.61)
PTQ: MSE-based range setting	7.68 (+0.60)
KD-QAT: Baseline (no freeze)	7.31 (+0.23)
KD-QAT: qkv-freeze (ours)	7.12 (+0.04)
KD-QAT: v-freeze (ours)	7.11 (+0.03)
KD-QAT: o-freeze (ours)	7.09 (+0.01)
KD-QAT: oqkv-freeze (ours)	6.99 (–0.09)
KD-QAT: ov-freeze (ours)	**6.98** (–0.10)

Table 1: Wikitext Perplexity for FP16 and 4-bit weight only quantized LLaMAv2-Chat Model using context length 2048 obtained through various PTQ and KD-QAT schemes. Perplexity (↓): the lower the better. Red color denotes >0.2 point and Green denotes <0.2 point perplexity change compared to FP16.

Finally, to enable on-device deployment of LLaMAv2-Chat model for practical uses, we evaluate our proposed ov-freeze W4A16 quantized model on multiple benchmarks. We use KD-QAT based weight quantized finetuned model, and quantize all activation outputs to INT16 using PTQ based min-max range setting to obtain the W4A16 model. As shown in Table 2, freezing ov-layers achieves competitive performance within 0.7% drop of FP16 model on average on CommonSense reasoning benchmarks, thereby improving significantly over both the best MSE based RTN and AdaRound baselines, both of which lose more than 2% accuracy compared to the FP16 network.

Model	PIQA(↑)	Arc-Easy(↑)	Arc-Challenge(↑)	Winogrande(↑)	OpenbookQA(↑)	Hellaswag(↑)	Average(↑)
FP16	76.66	69.65	44.37	66.38	43.8	75.42	62.71
PTQ							
MSE	75.79	65.53	41.47	66.22	41.00	72.76	60.46 (–2.25%)
AdaRound	74.59	66.62	40.7	66.54	40.60	72.28	60.22 (–2.49%)
KD-QAT							
no-freeze	76.66	67.38	42.75	**67.17**	41.80	74.14	61.65 (–1.06%)
ov-freeze	**77.26**	**69.11**	**42.83**	66.54	**42.20**	**74.15**	**62.02** (–0.69%)

Table 2: Evaluation of FP16 and various W4A16 Quantized LLaMAv2-Chat models using 1024 context length on Commonsense Reasoning benchmarks. (↑): the higher the better. Red color denotes >1% and Green denotes <1% loss of accuracy compared to FP16.

5 Conclusion

In this work, we proposed a light-weight quantization aware finetuning technique using knowledge distillation (KD-QAT) to improve the performance of popular W4A16 quantized LLM, LLaMAv2-Chat, for an important use case, on device chat applications, using commonly available datasets. We perform systematic study of the forward and backward pass of the LLaMAv2-Chat model, and analyze the output and gradient feature maps to hint towards extreme sensitivity of o- and v-layers in the multi-head attention modules. To improve the stability of the quantized finetuning, we proposed ov-freeze and experimented with popular 7B LLaMAv2-Chat model at 4-bit quantization level to show significant improvement in performance over vanilla QAT and achieve near float-point precision performance (within 0.7% drop of the FP16 model) on the CommonSense reasoning benchmarks.

q-투영 출력은 그림 3(a)에 표시됩니다. 이러한 전방 전달의 불안정성은 KD-QAT 동안 훈련 궤도를 쉽게 불안정하게 만들 수 있습니다. 그림 3은 은닉 모듈(레이어 ID) 10의 결과를 보여주지만, 다른 은닉 모듈에서도 유사한 경향을 관찰했음을 참고하십시오(부록 A 참조).

다음으로, 우리는 ov-freeze가 역전파에 미치는 영향을 경험적으로 탐구한다.

4.2.2 역방향 경로 분석

이제 마지막 섹션에서 다룬 동일한 숨은 모듈의 q-, k-, v-, o-투영 계층의 출력에서 기울기를 분석합니다. 그림 3(e-h)에 표시된 것처럼 기울기 노름 제곱 값을 분석합니다. 분명히, INT4 ov-동결(녹색 선)은 참조 FP16-동결 없음 훈련 실행(파란색 선)과 유사한 기울기 행동을 다시 보여줍니다. 반면, INT4 동결 없음 실험(빨간 선)은 참조에 비해 극적으로 다른 역전파 특성을 가지고 있습니다. 이러한 모든 불안정성은 최종 정확도의 저하와 불안정한 KD-QAT로 이어집니다.

4.3 결과

이 섹션에서는 제안된 동결 방식에 따라 FP16 및 양화 모델의 성능을 여러 벤치마크에서 비교합니다. 먼저 표 1은 Wikitext 데이터셋에서 PTQ 및 KD-QAT 최적화를 거친 4비트 가중치만 양화된 LLaMAv2-Chat의 퍼플렉시티를 보고합니다. 우리의 제거 분석에서는 o-, v-, ov-, qkv- 및 qkvo-와 같은 다양한 동결 방식을 고려하고 상대적으로 동결 계층을 관찰합니다.

Model	**Perplexity (ppl↓)**
FP16	7.08
PTQ: Min-Max range setting	9.69 (+2.61)
PTQ: MSE-based range setting	7.68 (+0.60)
KD-QAT: Baseline (no freeze)	7.31 (+0.23)
KD-QAT: qkv-freeze (ours)	7.12 (+0.04)
KD-QAT: v-freeze (ours)	7.11 (+0.03)
KD-QAT: o-freeze (ours)	7.09 (+0.01)
KD-QAT: oqkv-freeze (ours)	6.99 (–0.09)
KD-QAT: ov-freeze (ours)	**6.98** (–0.10)

표 1: FP16 및 4비트 가중치 전용 양화된 LLaMAv2-챗 모델의 위키텍스트 당혹도(context 길이 2048). 다양한 PTQ 및 KD-QAT 방식을 통해 얻었습니다. 당혹도(↓): 낮을수록 좋음. 빨간색은 >0.2 포인트, 초록색은 <0.2 포인트의 당혹도 변화를 나타냅니다.

더 높은 그래디언트 노름 값, 즉 o- 및 v-와 다른 레이어들은 KD-QAT 훈련의 안정성과 위키텍스트의 당혹감을 크게 향상시키는 데 도움이 되며, FP 모델과 동등하거나 더 나은 성능을 달성합니다. 표 1에서 볼 수 있듯이, ov- 레이어를 고정하는 것이 다른 모든 고정 방식과 PTQ 기준선을 능가하며, FP16 모델보다 더 나은 위키텍스트 당혹성을 달성합니다.

마지막으로, 실용적인 사용을 위한 LLaMAv2-Chat 모델의 온디바이스 배포를 가능하게 하기 위해, 우리는 여러 벤치마크에서 제안된 ov-freeze W4A16 정량화 모델을 평가합니다. 우리는 KD-QAT 기반 가중치 정량화 미세 조정 모델을 사용하고, PTQ 기반 최소-최대 범위 설정을 사용하여 모든 활성화 출력을 INT16으로 정량화하여 W4A16 모델을 얻습니다. 표 2에서 볼 수 있듯이, ov-레이어 동결은 FP16 모델에 비해 평균 0.7%의 성능 저하로 상식 추론 벤치마크에서 경쟁력 있는 성능을 달성하여, MSE 기반 RTN과 AdaRound 모두에 비해 크게 개선되었습니다. 두 기본 모델 모두 FP16 네트워크에 비해 정확도가 2% 이상 손실됩니다.

Model	**PIQA(↑)**	**Arc-Easy(↑)**	**Arc-Challenge(↑)**	**Winogrande(↑)**	**OpenbookQA(↑)**	**Hellaswag(↑)**	**Average(↑)**
FP16	76.66	69.65	44.37	66.38	43.8	75.42	62.71
PTQ							
MSE	75.79	65.53	41.47	66.22	41.00	72.76	60.46 (–2.25%)
AdaRound	74.59	66.62	40.7	66.54	40.60	72.28	60.22 (–2.49%)
KD-QAT							
no-freeze	76.66	67.38	42.75	**67.17**	41.80	74.14	61.65 (–1.06%)
ov-freeze	**77.26**	**69.11**	**42.83**	66.54	**42.20**	**74.15**	**62.02** (–0.69%)

표 2: Commonsense Reasoning 벤치마크에서 1024 컨텍스트 길이를 사용하여 FP16 및 다양한 W4A16 정량화된 LLaMAv2-Chat 모델의 평가. (↑): 값이 높을수록 좋음. 빨간색은 >1% 및 초록색은 FP16에 비해 < 1%의 정확도 손실을 나타냄.

5 결론

이 연구에서 저자는 지식 증류(KD-QAT)를 활용한 경량 양자화 인식 미세 조정 기술을 제안하여 일반적인 W4A16 양자화 LLM인 LLaMAv2-Chat의 성능을 개선하고, 일반적으로 이용 가능한 데이터셋을 사용하여 중요한 사용 사례인 온디바이스 채팅 애플리케이션에 적용했습니다. LLaMAv2-Chat 모델의 순전파 및 역전파를 체계적으로 연구하고, 출력 및 기울기 특징 맵을 분석하여 다중 헤드 어텐션 모듈의 o-층 및 v-층의 극심한 민감도를 시사했습니다. 양자화 미세 조정의 안정성을 개선하기 위해 ov-동결을 제안하고, 인기 있는 7B LLaMAv2-Chat 모델을 4비트 양자화 수준에서 실험하여 바닐라 QAT에 비해 성능이 크게 향상되고, FP16 모델의 0.7% 이내의 성능 저하로 커먼센스 추론 벤치마크에서 부동 소수점 정확도 성능에 근접함을 보였습니다.

REFERENCES

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 696–697, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

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. *arXiv preprint arXiv:2305.17888*, 2023.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1325–1334, 2019.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pp. 7197–7206. PMLR, 2020.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. *arXiv preprint arXiv:2203.11086*, 2022.

Sangeetha Siddegowda, Marios Fournarakis, Markus Nagel, Tijmen Blankevoort, Chirag Patel, and Abhijit Khobare. Neural network quantization with ai model efficiency toolkit (aimet). *arXiv preprint arXiv:2201.08442*, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=f-fVCElZ-G1.

참조: 야시 발가트, 이진원 리, 마르쿠스 나겔, 티멘 블랑코보르트, 노준 곽. Lsq+: 학습 가능한 오프셋과 더 나은 초기화를 통한 저비트 양자를 개선합니다. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*에서, pp. 696-697, 2020. 프라풀라 다리와알, 알렉산더 니콜. 확산 모델은 GAN을 이미지 합성에 능가합니다. *Advances in neural information processing systems*, 34:8780-8794, 2021. 진 동, 저웨이 야오, 아미르 그홀라미, 마이클 W 마호니, 커트 케저. 하크: 혼합 정밀도를 가진 신경망의 헤시안 인식 양자화. *Proceedings of the IEEE/CVF International Conference on Computer Vision*에서, pp. 293-302, 2019. 스티븐 K 에서, 제프리 L 맥킨스트리, 딥카 바블라니, 라티나쿠마르 아푸스와미, 다르멘드라 S 모다. 학습된 단계 크기 양자화. *arXiv preprint arXiv:1902.08153*, 2019. 엘리아스 프란타르, 살레 아슈크부스, 토르스텐 호플러, 단 알리스타르. GPTQ: 생성 사전 훈련 변환기에 대한 정확한 사후 훈련 양자화, 2023. 조나단 호, 아자이 자인, 피터 아벨. 소음 제거 확산 확률 모델. *Advances in neural information processing systems*, 33:6840-6851, 2020. 베네트 제이콥, 스커르만트 클리그스, 보 첸, 멩롱 주, 매튜 탕, 앤드류 하워드, 하트비그 아담, 드미트리 칼레니첸코. 효율적인 정수 산술 전용 추론을 위한 신경망 양자화 및 훈련.
*Proceedings of the IEEE conference on computer vision and pattern recognition*에서, pp. 2704-2713, 2018. 라그후라마 크리슈나모오티. 효율적인 추론을 위한 깊은 합성곱 신경망 양자화: 백서. *arXiv preprint arXiv:1806.08342*, 2018. 제춘 류, 바르라스 오구즈, 창셍 자오, 어니 창, 피에르 스톡, 야샤르 메흐다드, 양양 시, 라그후라마 크리슈나모오티, 비카스 찬드라. LLM-QAT: 데이터 무료 양자화 인식 훈련을 위한 대규모 언어 모델. *arXiv preprint arXiv:2305.17888*, 2023. 마르쿠스 나겔, 마르트 반 바알렌, 티멘 블랑코보르트, 맥스 웰링. 데이터 무료 양자화를 위한 가중치 동등화 및 편향 수정. *Proceedings of the IEEE/CVF International Conference on Computer Vision*에서, pp. 1325-1334, 2019. 마르쿠스 나겔, 라나 알리 암자드, 마르트 반 바알렌, 크리스토스 루이즈, 티멘 블랑코보르트. 위나 아래? 사후 훈련 양자화를 위한 적응형 반올림. *International Conference on Machine Learning*에서, pp. 7197-7206. PMLR, 2020. 마르쿠스 나겔, 마리오스 포우르나라키스, 라나 알리 암자드, 예르세이 본다렌코, 마르트 반 바알렌, 티멘 블랑코보르트. 신경망 양자화에 대한 백서. *arXiv preprint arXiv:2106.08295*, 2021. 마르쿠스 나겔, 마리오스 포우르나라키스, 예르세이 본다렌코, 티멘 블랑코보르트. 양자화 인식 훈련에서 진동 극복. *arXiv preprint arXiv:2203.11086*, 2022. 산게타 시데고우다, 마리오스 포우르나라키스, 마르쿠스 나겔, 티멘 블랑코보르트, 치라그 파텔, 아브히짓 코바레. AI 모델 효율성 도구를 사용한 신경망 양자화. *arXiv preprint arXiv:2201.08442*, 2022. 우고 투브론, 루이 마틴, 케빈 스톤, 피터 알버트, 아자드 알마하이리, 야스민 바바에, 니콜라이 바슬리코프, 수미야 바트라, 프라즈왈 바르가바, 슈루티 보살레 외. 라마 2: 오픈 재단 및 미세 조정된 채팅 모델. *arXiv preprint arXiv:2307.09288*, 2023. 저웨이 야오, 레자 야즈다니 아미나바디, 민지아 장, 샤오샤 우, 콩롱 리, 유시옹 헤. 제로퀀트: 효율적이고 저렴한 대규모 변환기를 위한 사후 훈련 양자화. 앨리스 H. 오, 알레크 아가왈, 다니엘 벨그레이브, 쿤현 초 (편집). *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum? id=f-fVCElZ-G1.

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.

A Additional Results

Forward and Backward Signal Propagation Results on Other LLaMAv2-Chat Modules. Fig. 4, 5, and 6 show that our observations for hidden module 10 (i.e., Layer ID: 10 in Fig. 3) hold across other hidden modules. Clearly, the no-freeze mean forward pass for v-projection layer is unstable for all the hidden modules shown in Fig. 4, 5, and 6. The proposed method ov-freeze makes the forward and backward pass characteristics similar to those seen during floating point training.

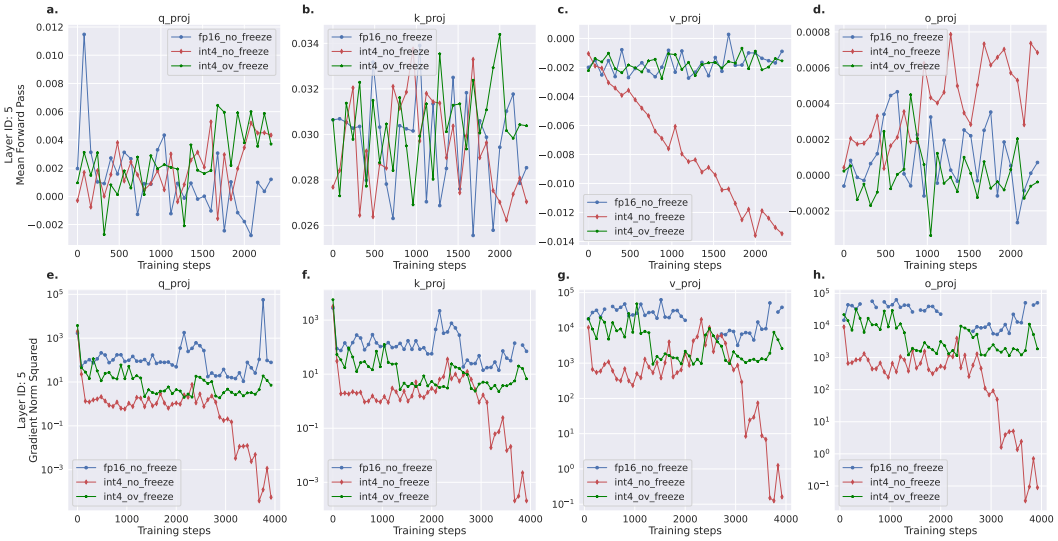

Figure 4: Layer ID: 5. Forward and backward pass signal propagation analysis on 5th hidden LLaMAV2-chat self-attention module. Our proposed ov-freeze makes the quantized model's forward pass and gradients more similar to those observed during FP16 training.

장수잔, 롤러 스티븐, 고얄 나만, 아르텍스 미켈, 첸 모야, 첸 수오후이, 데완 크리스토퍼, 디아브 모나, 리 시안, 린 빅토리아 시, 미하이토르 토도르, 오트 마일, 슬레이퍼 샘, 슈스터 커트, 시미그 다니엘, 쿠라 싱 푸닛, 스리다르 안잘리, 왕 티안루, 제틀레마이 루크. Opt: 오픈 사전 학습된 트랜스포머 언어 모델, 2022.

추가 결과

다른 LLaMAv2-Chat 모듈에서의 전방 및 후방 신호 전파 결과. 그림 4, 5 및 6은 숨겨진 모듈 10 (즉, 그림 3의 레이어 ID: 10)에 대한 우리의 관찰이 다른 숨겨진 모듈에도 적용됨을 보여줍니다. 분명히, 그림 4, 5 및 6에 표시된 모든 숨겨진 모듈에 대해 {v*} 투영 레이어의 비동결 평균 전방 패스는 불안정합니다. 제안된 동결 방법은 부동 소수점 학습 중에 관찰된 전방 및 후방 패스 특성과 유사하게 만듭니다.

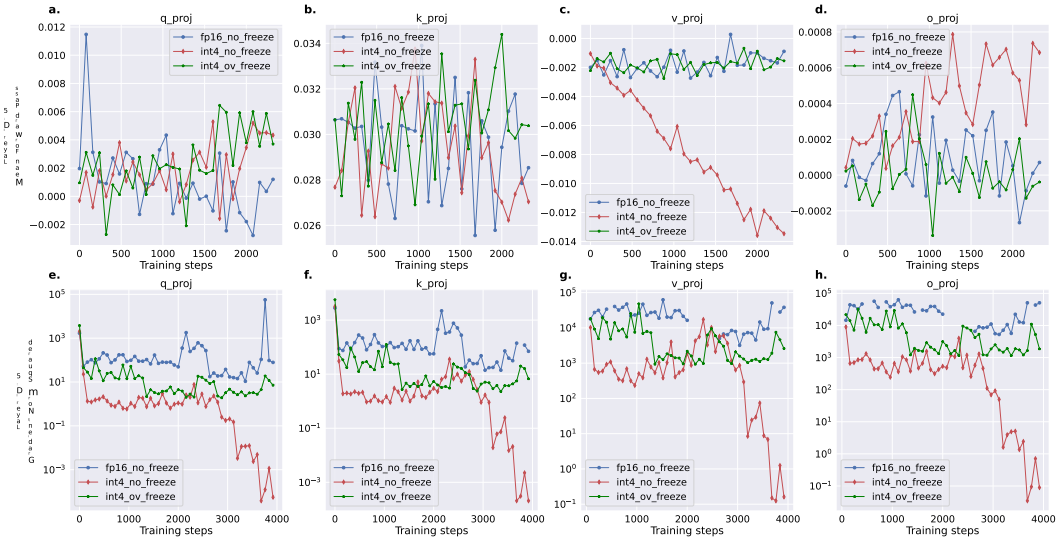

그림 4: 계층 ID: 5. 5번째 숨겨진 LLaMAV2-chat 자기 주의 모듈에 대한 전방 및 후방 패스 신호 전파 분석. 우리가 제안한 ov-freeze는 양화 모델의 전방 패스와 기울기가 FP16 훈련 중에 관찰된 것과 더 유사하게 만듭니다.

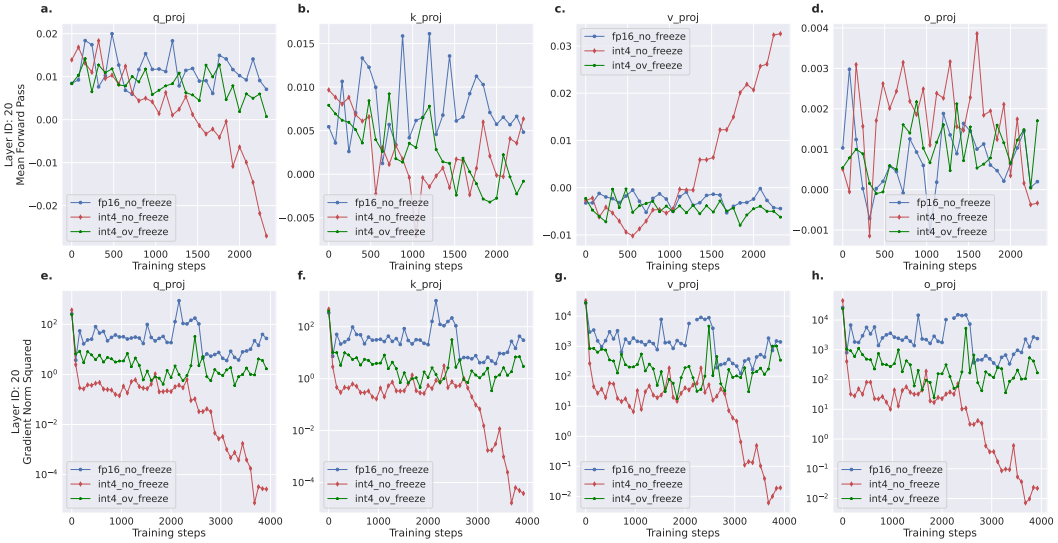

Figure 5: Layer ID: 20. Forward and backward pass signal propagation analysis on 20th hidden LLaMAV2-chat self-attention module. Our proposed ov-freeze makes the quantized model's forward pass and gradients more similar to those observed during FP16 training.

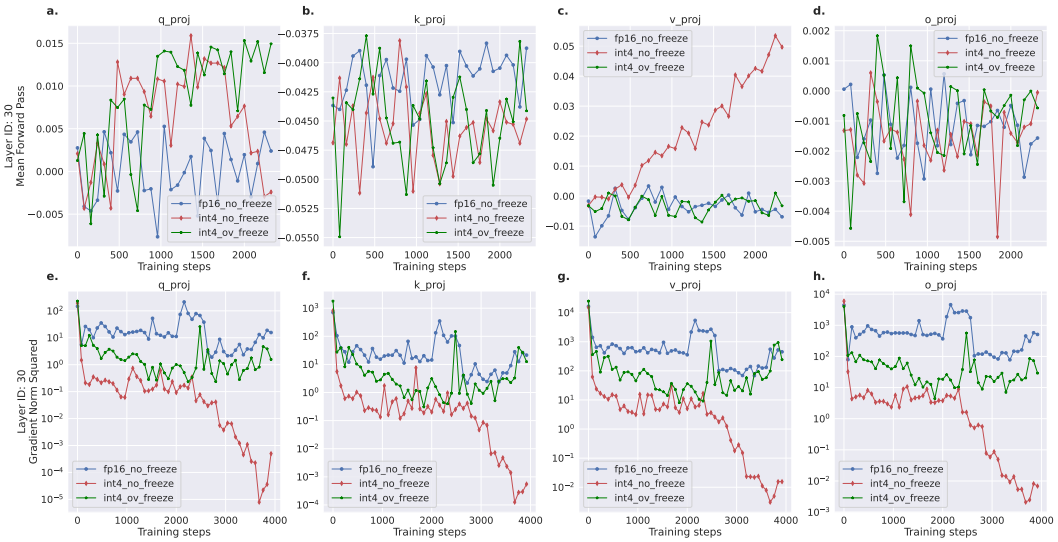

Figure 6: Layer ID: 30. Forward and backward pass signal propagation analysis on 30th hidden LLaMAV2-chat self-attention module. Our proposed ov-freeze makes the quantized model's forward pass and gradients more similar to those observed during FP16 training.

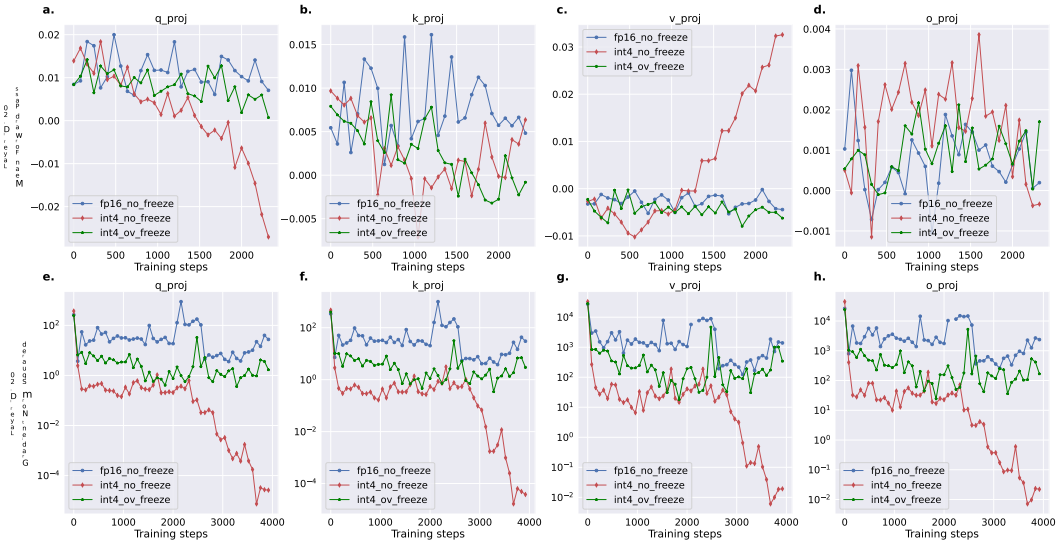

그림 5: 계층 ID: 20. 20번째 숨겨진 LLaMAV2-chat 자기 주의 모듈에 대한 전방 및 후방 패스 신호 전파 분석. 우리가 제안한 ov-freeze는 정량화된 모델의 전방 패스와 기울기가 FP16 훈련 중에 관찰된 것과 더 유사해집니다.

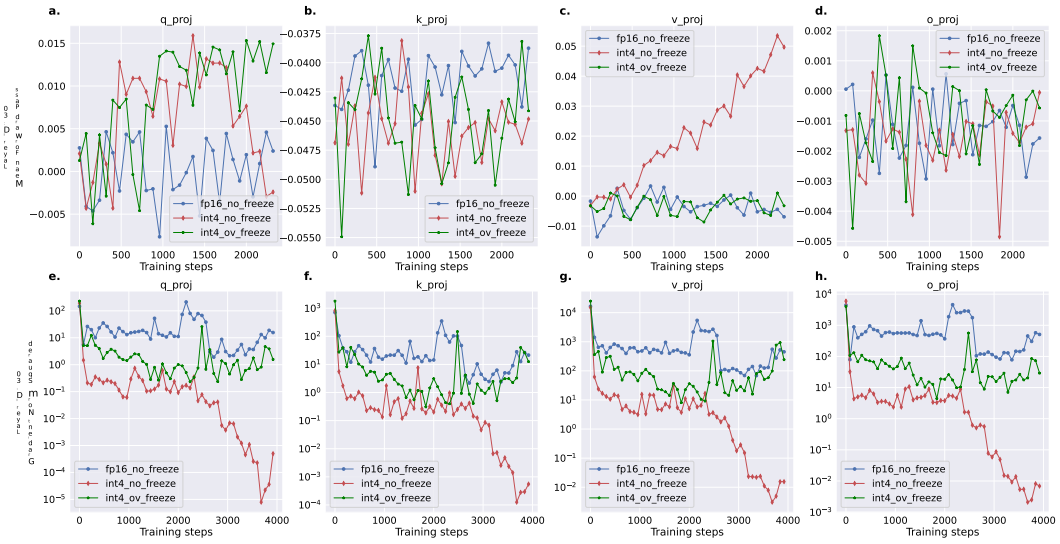

그림 6: 레이어 ID: 30. 30번째 숨겨진 LLaMAV2-chat 자기 주의 모듈에 대한 전방 및 후방 패스 신호 전파 분석. 우리가 제안한 ov-freeze는 정량화된 모델의 전방 패스와 기울기가 FP16 훈련 중에 관찰된 것과 더 유사해집니다.