양자화를 이용한 Diffusion 모델 경량화

서용석, 임동혁
한국전자통신연구원(ETRI)

yongseok@etri.re.kr

Lightweight Diffusion Models via Quantization

Yongseok Seo, Dong-Hyuck Im
Electronics and Telecommunications Research Institute

요 약

본 논문은 Diffusion 모델의 가중치를 양자화하여 이미지 생성 속도를 가속화하고 요구되는 VRAM 을 감소시킬수 있는 경량화 방법을 제안한다. 최근의 이미지 생성을 위한 Diffusion 모델들은 연산량이 매우 커서 고성능의 GPU 컴퓨팅 자원이 없는 환경에서는 실행이 어려운데, 제안한 양자화 기법을 활용하면 저사양의 환경에서도 경량화된 모델을 활용하여 효과적인 고품질 이미지 생성이 가능하다. 실험결과, 제안된 양자화 기반 모델 경량화 방법은 생성 이미지 품질을 유지하면서 생성속도 개선과 함께 GPU 메모리 요구량을 크게 개선하였다.

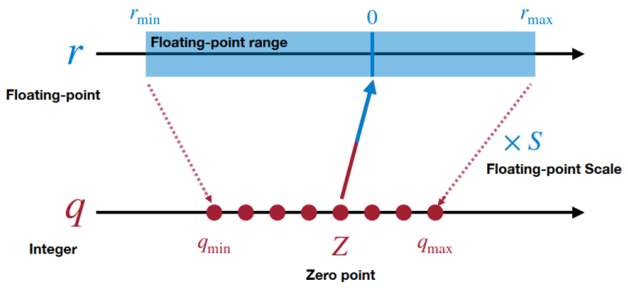
Ⅰ. 서론

Diffusion 모델은 최근 인공지능 분야에서 가장 주목받는 생성 모델 중 하나로 이미지 생성과 텍스트-이미지 변환과 같은 다양한 작업에서 뛰어난 성능을 입증하고 있다. Diffusion 은 데이터에 점진적으로 노이즈를 추가한 뒤, 이를 역으로 제거하는 과정을 학습함으로써 새로운 데이터를 생성하는데, 이러한 방식은 기존 생성모델 대비 높은 안정성과 우수한 일반화 성능을 제공한다. 그러나 Diffusion 모델은 대규모 파라미터를 포함하여 연산량이 매우 크기 때문에, 고성능 컴퓨팅 자원이 없는 환경에서는 실행이 어려운 경우가 많다. 이에 따라 최근에는 이를 효율적으로 경량화하는 연구가 활발히 진행되고 있으며, 이 중에서도 양자화(Quantization) 기법은 핵심적인 해결방안으로 주목받고 있다.

본 논문에서는 Stable Diffusion 과 같은 Latent Diffusion 모델의 가중치를 저비트로 양자화함으로써 모델의 경량화를 도모하고, 이미지 생성에 요구되는 시간과 메모리를 절감하는 방법을 제안한다. 또한, 제안된 방법을 SDXL-Lightning 모델[1]에 적용하여 경량화된 모델의 이미지 생성 시간, GPU 메모리 사용량 및 생성 이미지의 품질을 종합적으로 분석한다.

Ⅱ. 본론

딥러닝 모델의 양자화는 연산 효율성을 높이기 위해 모델의 가중치나 활성화 값을 부동소수점에서 더 낮은 비트의 정수형 값으로 변환하여 연산량과 메모리 사용량을 줄여주는 기법이다. 이때 부동소수점 값을 정수로 매핑하기 위해 단순 반올림을 사용하면 오차가 매우 커지므로, 모델의 각 채널 또는 레이어에서 데이터 분포에 따라 양자화 파라미터를 결정한다.



$$scale = \frac{r_{max} - r_{min}}{q_{max} - q_{min}} \qquad zero\_point = round(-\frac{r_{min}}{scale})$$

$$\boldsymbol{Q\_value = round(\frac{R\_value}{scale}) + zero\_point)}$$

그림 1. 선형 양자화 방법

양자화는 FP 형 데이터를 하나의 scale factor 로 나눠주고 Rounding 과 Clipping 을 함으로써 INT 형의 양자화된 구간으로 매핑을 한다. 이때 사용되는 zero-point 는 양자화 전 0 값이 양자화된 구간으로 매핑되었을때의 값을 의미하고, scale factor 와 zero-point 를 양자화 파라미터라 하며 모델의 가중치와 활성화 값의 분포를 통해 결정된다.

양자화된 모델은 기존의 소수점 연산에 비해 메모리 대역폭이 감소하며, 연산 속도가 향상되어 전력 소모가 큰 폭으로 줄어든다. Diffusion 모델은 기본적으로 연속적인 변환 과정을 거치므로 이러한 양자화의 장점이 더 크게 부각된다. 그러나 양자화된 모델은 정밀도가 낮아지는 과정에서 발생하는 손실 때문에 생성 품질 손실을 초래할 수 있으며, 이러한 성능저하를 최소화하기 위한 다양한 연구가 진행되고 있다 [2].

뉴럴넷에 대해 양자화를 진행하는 이유 중 하나는 메모리 사용량을 줄이기 위해서인데, 대부분의 뉴럴넷은 파라미터 수가 과잉상태이기 때문에 양자화를 통한

경량화 여지가 많다. SDXL FP32 모델의 추론에는 약 12~24GB 의 GPU 메모리 용량이 필요하며 이는 일반 소비자용 GPU 의 메모리 용량을 상회하는 수준이다. 일반적인 소비자용 GPU 는 10GB 이하의 메모리를 가지며, 이러한 제한된 환경에서는 모델 추론 시 CPU/NVMe 오프로딩이 발생하여 매우 느린 속도를 갖게 된다. 따라서 양자화를 통해 메모리 사용량을 줄이면 제한된 환경에서도 속도 저하 없이 빠른 추론이 가능해진다.

대부분의 이미지 생성용 Diffusion 모델들(Stable Diffusion, Imagen)은 latent feature 를 다운샘플링/업샘플링하는 디노이징 백본으로 UNet 을 채택하고 있는데 본 논문에서는 Stable Diffusion 을 지식증류하여 가속화한 SDXL-Lightning 모델을 기반으로 양자화 연구를 수행하였다.

LLM 계열에서는 양자화로 인한 정밀도 손실이 품질 저하로 이어지지 않도록 활성화와 가중치 값의 분포를 부드럽게 해주는 스무딩 기법을 사용한다. SmoothQuant [3]에서는 이상치값을 스무딩하기 위해 smoothing factor 를 정의하여, 각 활성화 값을 이 factor 로 나누고 가중치에 같은 값을 곱해줌으로써 수학적 동등성은 유지하면서 활성화 값이 좀더 균일한 분포가 되어 양자화 효율성을 높일 수 있다고 한다.

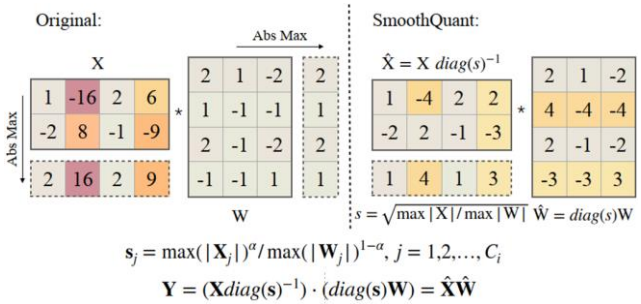


그림 2. SmoothQuant 주요 아이디어

아래 그림처럼 SmoothQuant 이후 활성화의 이상치가 가중치로 이전되어 가중치의 분포가 약간 불균일해지나 전체적으로는 활성화와 가중치 모두 양자화하기 쉬운 분포를 갖게된다.

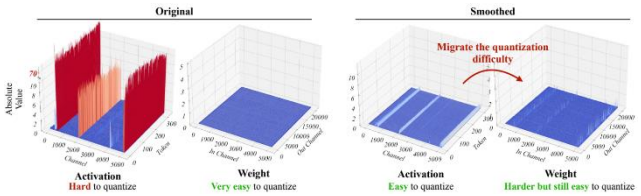


그림 3. SmoothQuant 전과 후 활성화와 가중치 분포

하지만 SmoothQuant 기법을 Stable Diffusion 모델에 그대로 적용하면 노이즈 형태의 이미지가 출력된다. Diffusion 모델의 각 레이어를 분석한 결과 활성화 분포가 LLM 모델에 비해 더 균일하지 않은 것이 원인으로 파악된다. 이러한 상황에서 활성화 값의 max vector 까지 나눠주는 연산을 적용하면 성능이 더 저하되기 때문이다. 따라서 가중치의 max vector 만으로 스무딩을 하고, 수학적 동등성 유지를 위해 활성화에 같은 값을 곱해주는 방식을 시도하였다. 즉 UNet 의 모든 선형 레이어 가중치를 2 차원 행렬 형태에서 스무딩을 적용하여 양자화에 유리한 범위로 변경한 후, 아래와 같은 동적 양자화 과정을 진행한다. *(X: activation, W: weight, Q: Quantization, D: diag(s))*

$$Y = X \cdot (Q(WD)D^{-1})$$

본 논문에서는 SDXL 기본모델을 점진적 적대적 확산 증류 기법으로 이미지 생성 속도를 향상시킨 SDXL-Lightning 4 step 모델을 대상으로 양자화 과정을 추가 적용하여 이미지 품질 저하없이 보다 빠른 추론 속도와 메모리 절감 효과를 확인할 수 있었다.

표 1. 이미지 생성 시간/메모리 비교 (생성 이미지 1024x1024)

UNet	Offloading	RTX6000		RTX3060	
		Time	VRAM	Time	VRAM
FP32	Offloading	5.54	10225	64.68	10229
INT8	No Offloading	1.67	6690	4.68	6693

(a) 양자화 미적용 (b) 양자화 적용

그림 4. 모델 양자화 전(위)/후(아래) 생성된 이미지 품질

Ⅲ. 결론

본 논문에서는 양자화를 이용한 Diffusion 모델 경량화 방법을 제안하였다. SDXL-Lightning 모델의 UNet 가중치 값을 8 비트로 양자화 적용하여 이미지 품질 저하는 최소화하면서 생성 시간 단축과 함께 GPU 메모리 사용량을 대폭 절감할 수 있음을 확인하였다.

ACKNOWLEDGMENT

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2025 년도 문화체육관광 연구개발사업 (연구개발과제명: 공연 콘텐츠의 고해상도(8K/16K) 서비스를 위한 AI 기반 영상확장 및 서비스 기술개발, 연구개발과제번호: RS-2024-00395886)

참 고 문 헌

[1] S. Lin, A. Wang, X. Yang, “SDXL-Lightning: Progressive Adversarial Diffusion Distillation.” arXiv:2402.13929, 2024.

[2] 임동혁, 서용석, “Stable Diffusion 모델의 가속화 및 경량화 기법.” 한국통신학회추계학술발표회, 2024.

[3] G. Xiao, J, Lin, M. Seznec, H. Wu, J. Demouth, S. Han, “SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models.” in Proc. ICML 2023.