Quantization

2026. 03. 13

Korea University

Data **M**ining & **Q**uality **A**nalytics Lab.

최지형

Data Mining
Quality Analytics

발표자 소개

❖ **최지형 (Jihyung Choi)**

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S Student (2024.09 ~ Present)

❖ **Research Interest**

- On-Device AI
- Fine-Tuning Foundation Models
- Agent AI

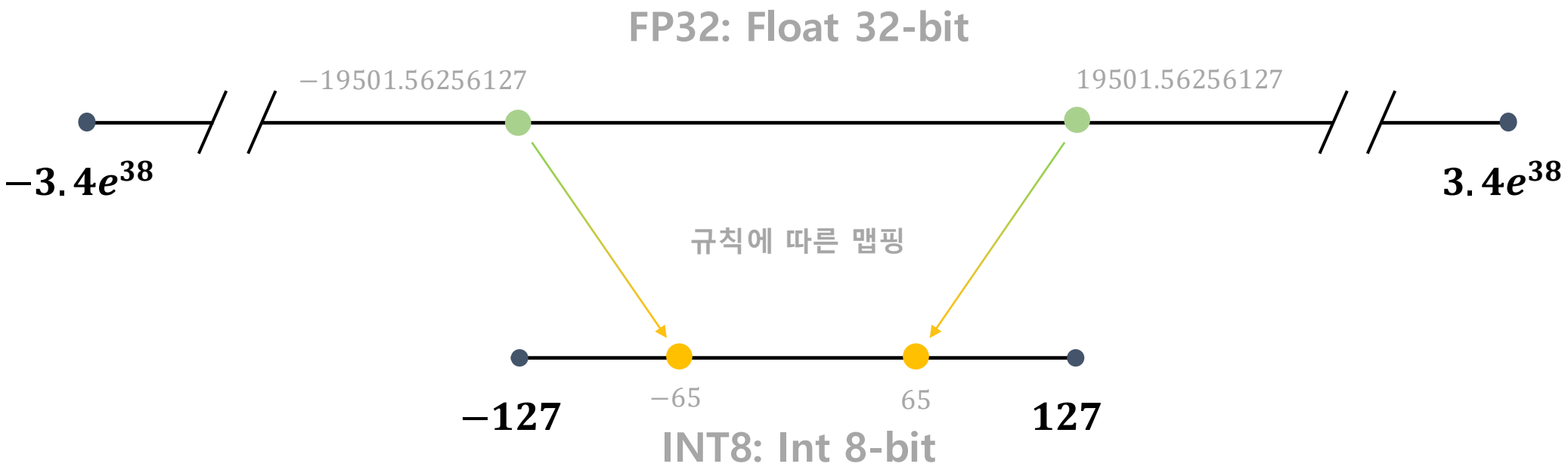
❖ **Contact**

- jibro@korea.ac.kr

Basic of Quantization

❖ Quantization

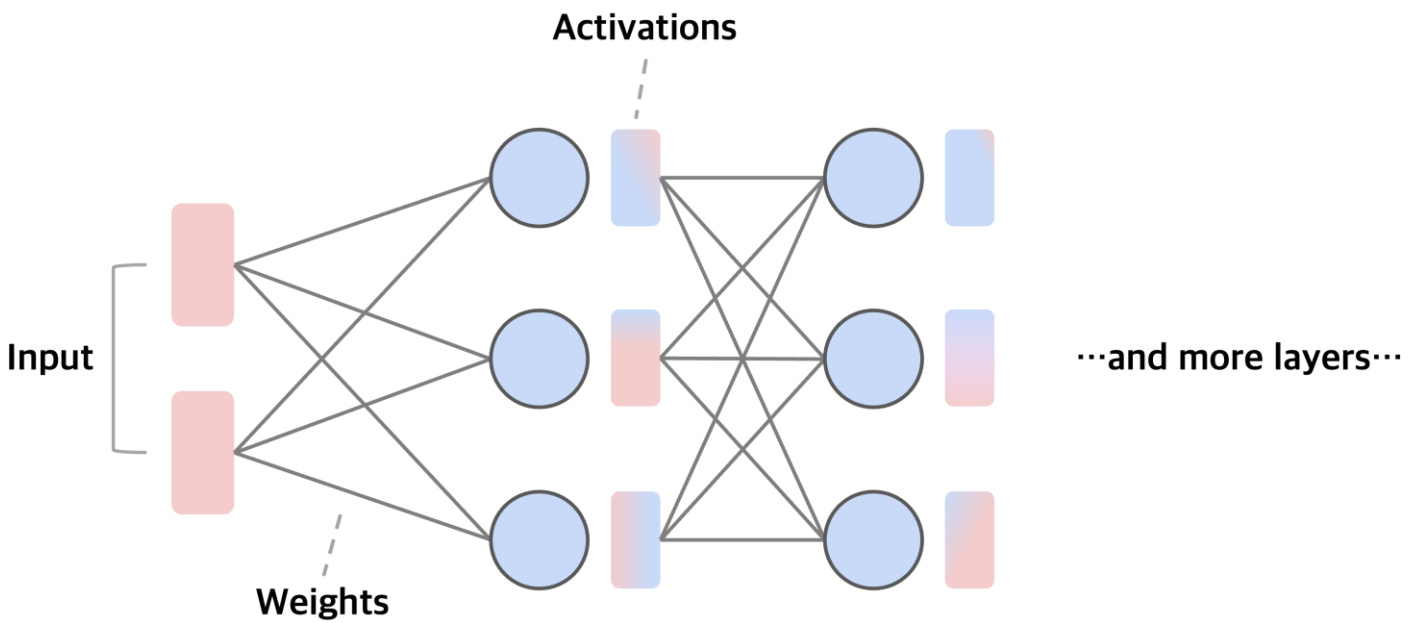
- ➢ 실수를 정수로 맵핑 (e.g. FP32 → Int8)
- ➢ 추론 시, weight와 activation 값에 quantization 적용
- ➢ 효과: 모델 사이즈 감소와 추론 속도 개선



Basic of Quantization

❖ Quantization

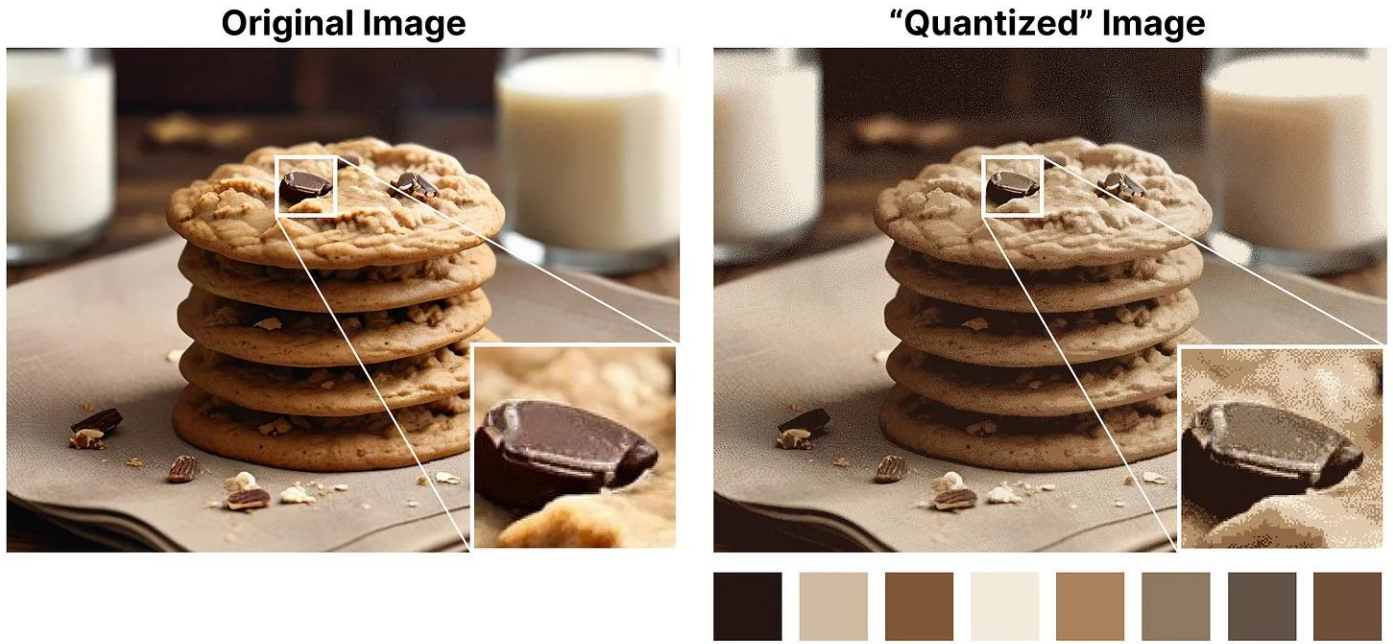
- ➢ 실수를 정수로 맵핑 (e.g. FP32 → Int8)
- ➢ 추론 시, weight와 activation 값에 quantization 적용
- ➢ 효과: 모델 사이즈 감소와 추론 속도 개선



Basic of Quantization

❖ Quantization

- ➢ 실수를 정수로 맵핑 (e.g. FP32 → Int8)
- ➢ 추론 시, weight와 activation 값에 quantization 적용
- ➢ 효과: 모델 사이즈 감소와 추론 속도 개선



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization

Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

변환하고자 하는 실수

Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

결정해야 하는 파라미터

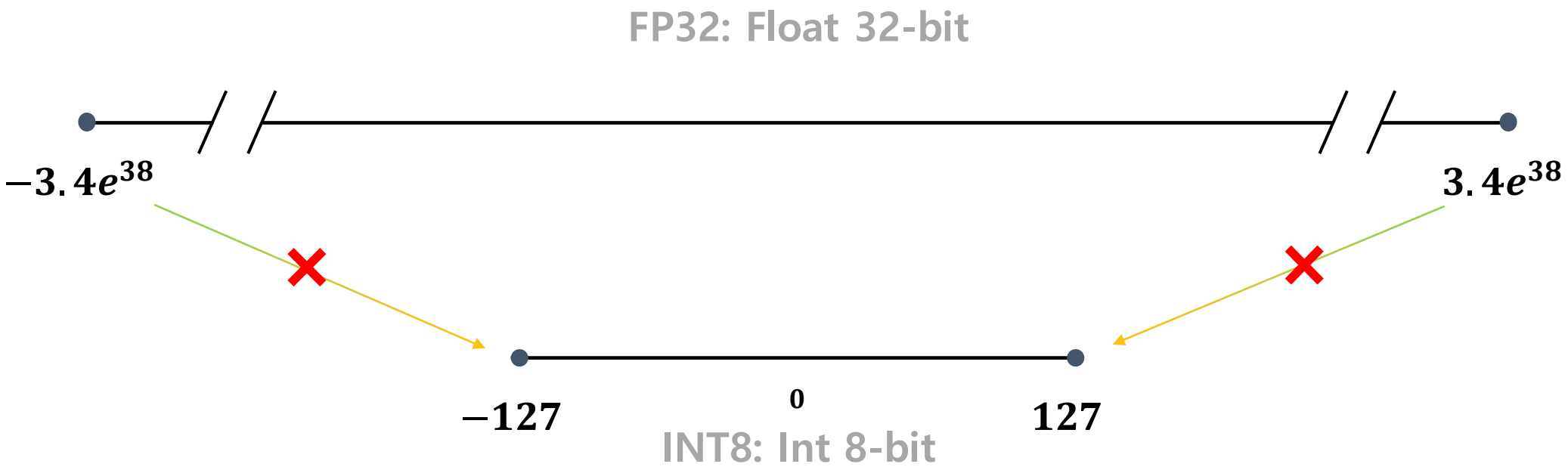
Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

결정해야 하는 파라미터



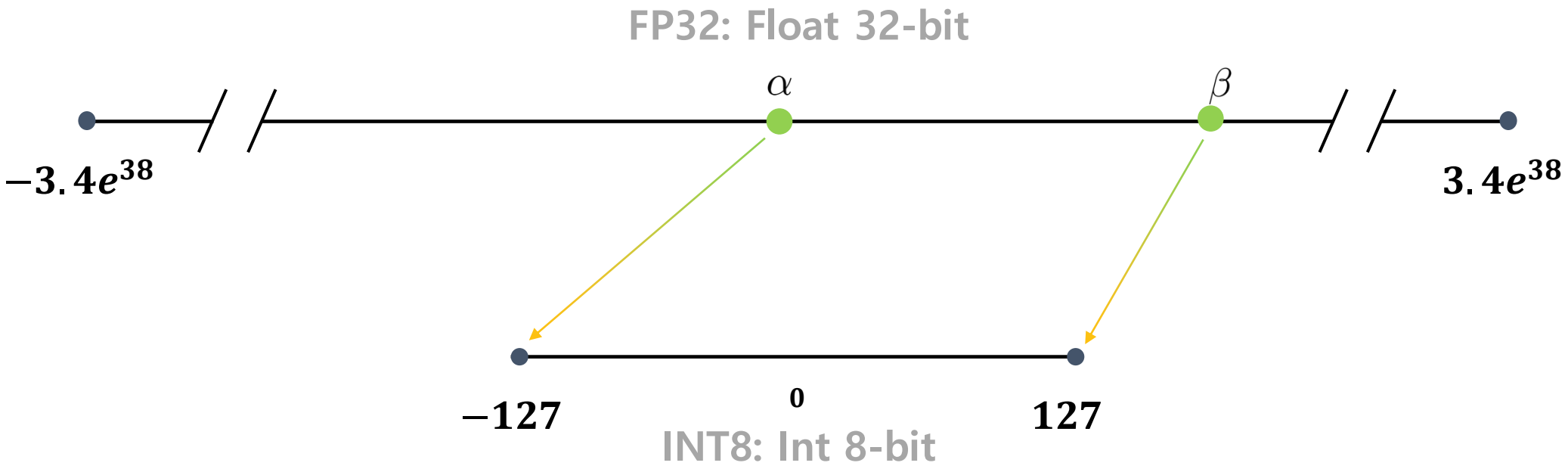
Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

결정해야 하는 파라미터



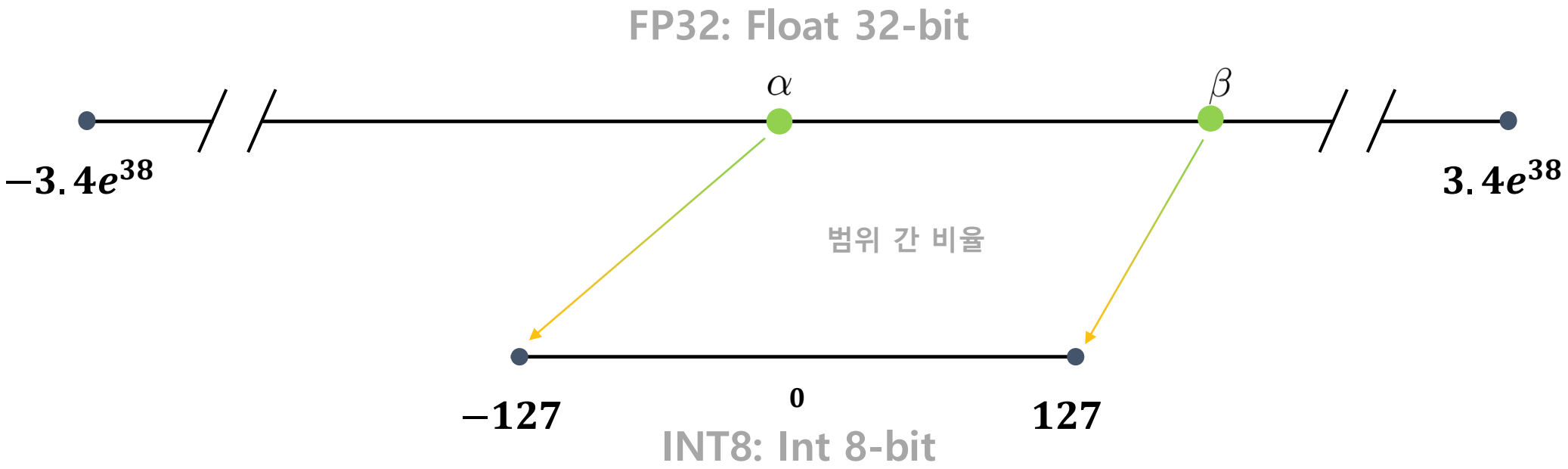
Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{\beta - \alpha}{254}$$

Scaling Factor

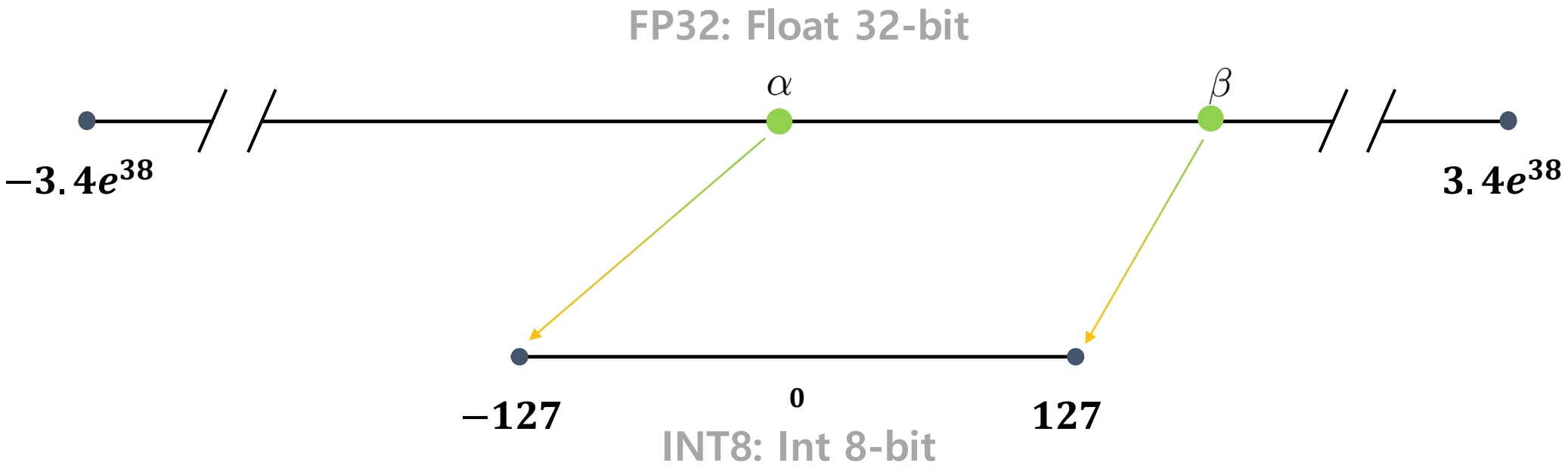


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{\beta - \alpha}{254}$$

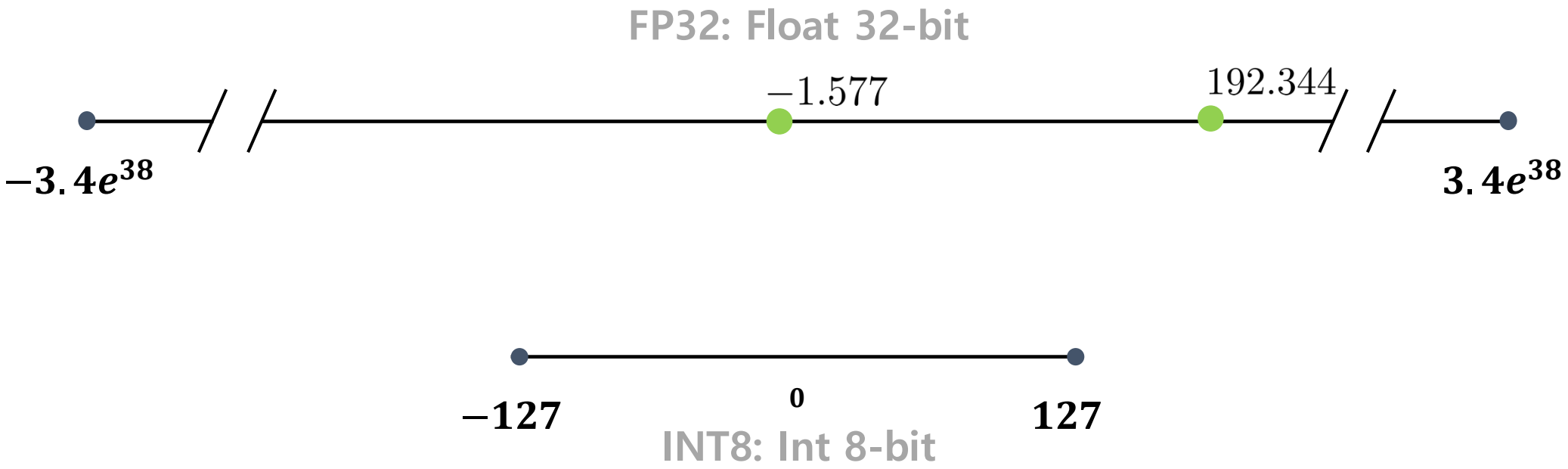


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{\beta - \alpha}{254}$$

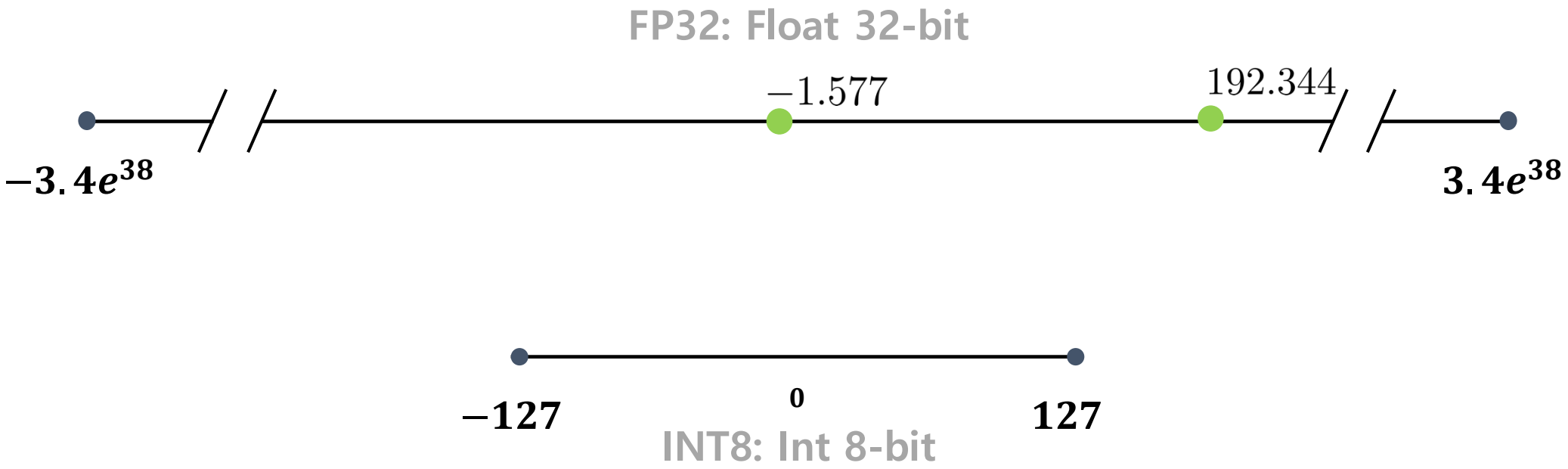


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{192.344 - (-1.577)}{254} = 0.763$$

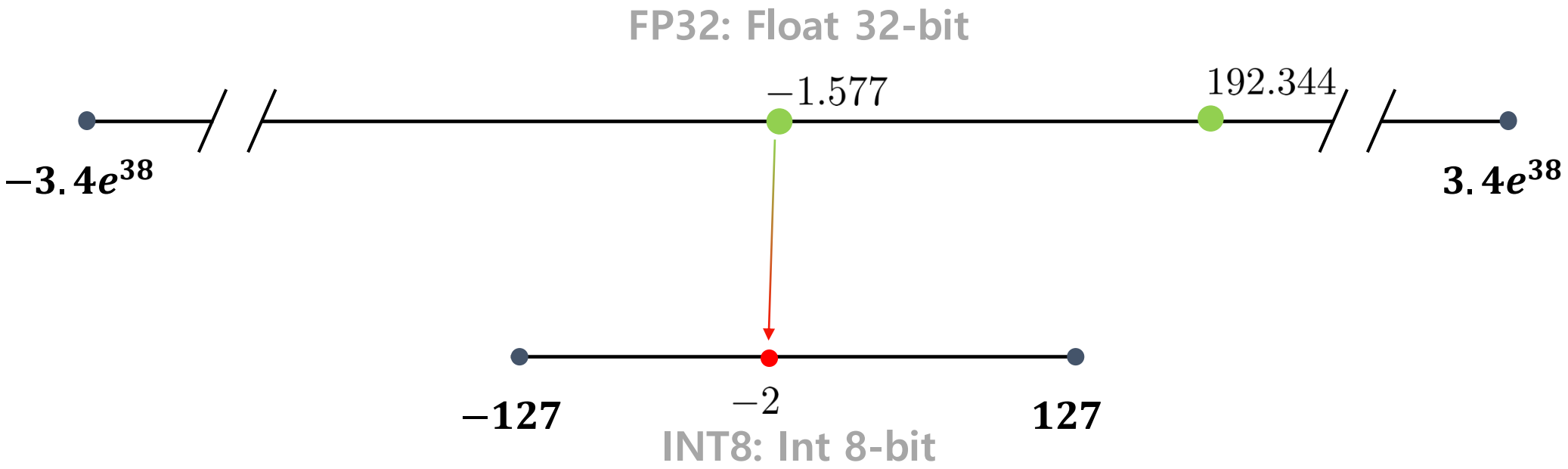


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{192.344 - (-1.577)}{254} = 0.763$$

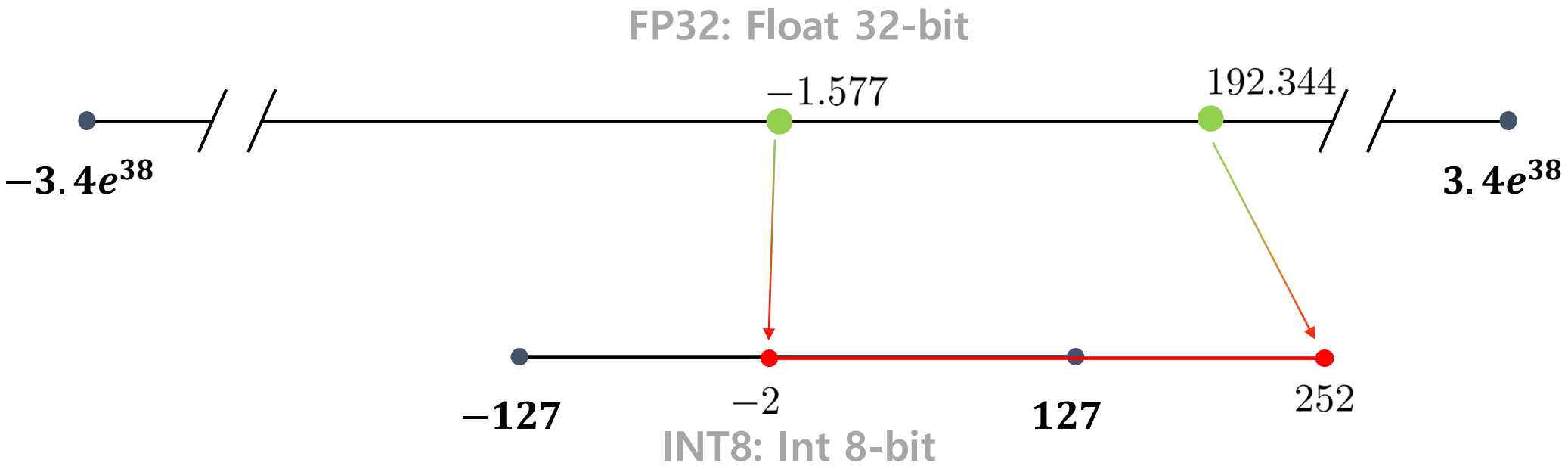


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{192.344 - (-1.577)}{254} = 0.763$$

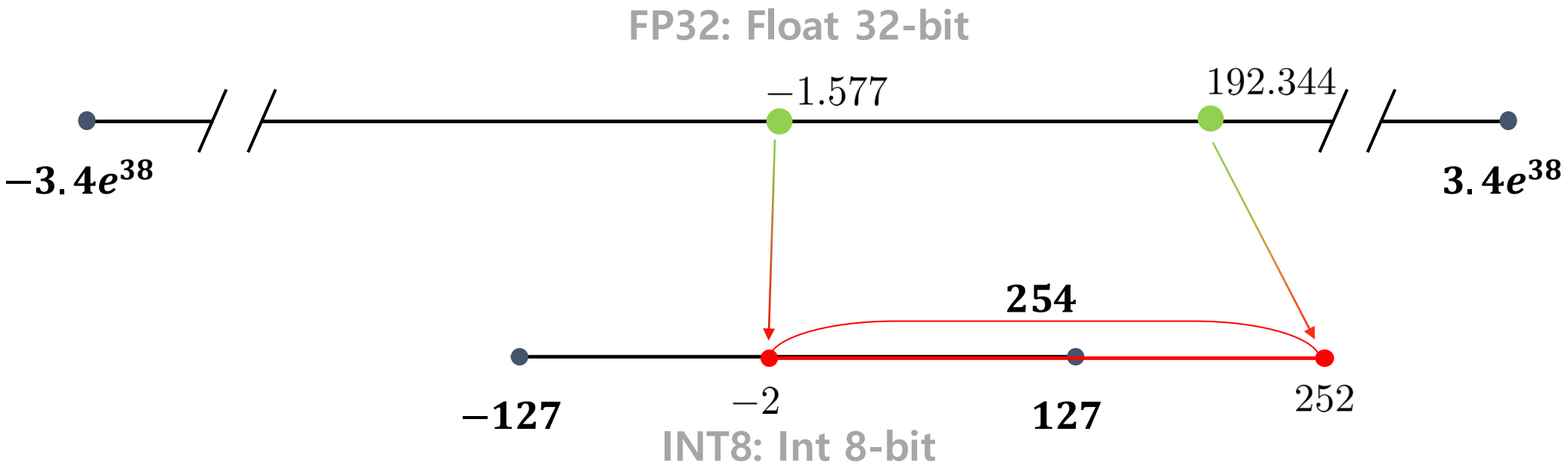


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{192.344 - (-1.577)}{254} = 0.763$$

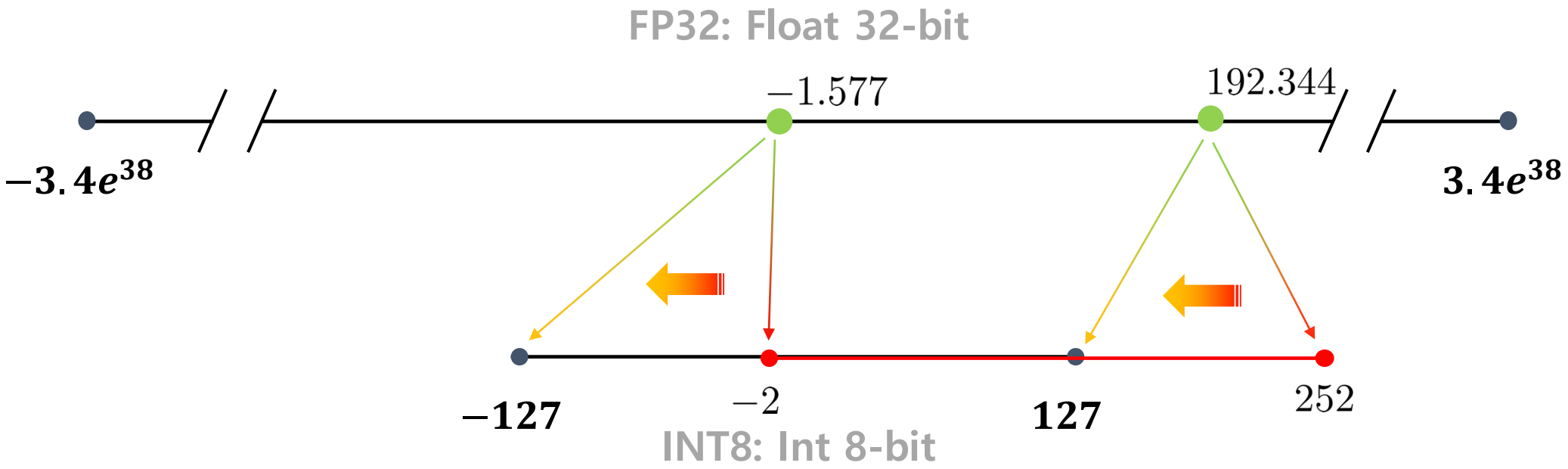


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z) \qquad s = \frac{192.344 - (-1.577)}{254} = 0.763$$



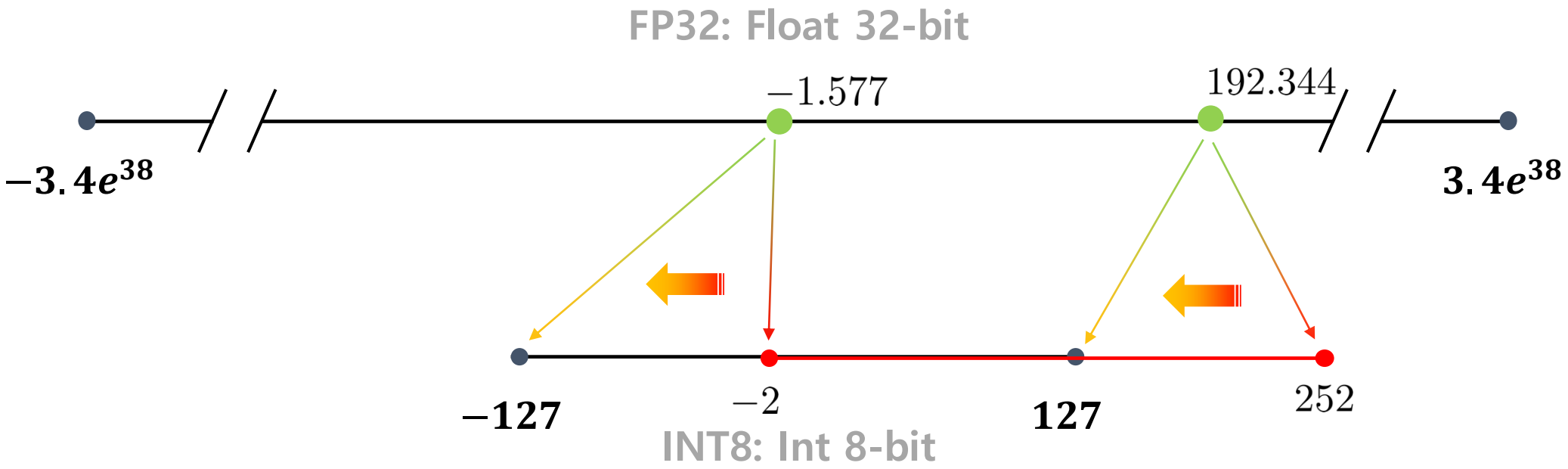
Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

Zero-point integer



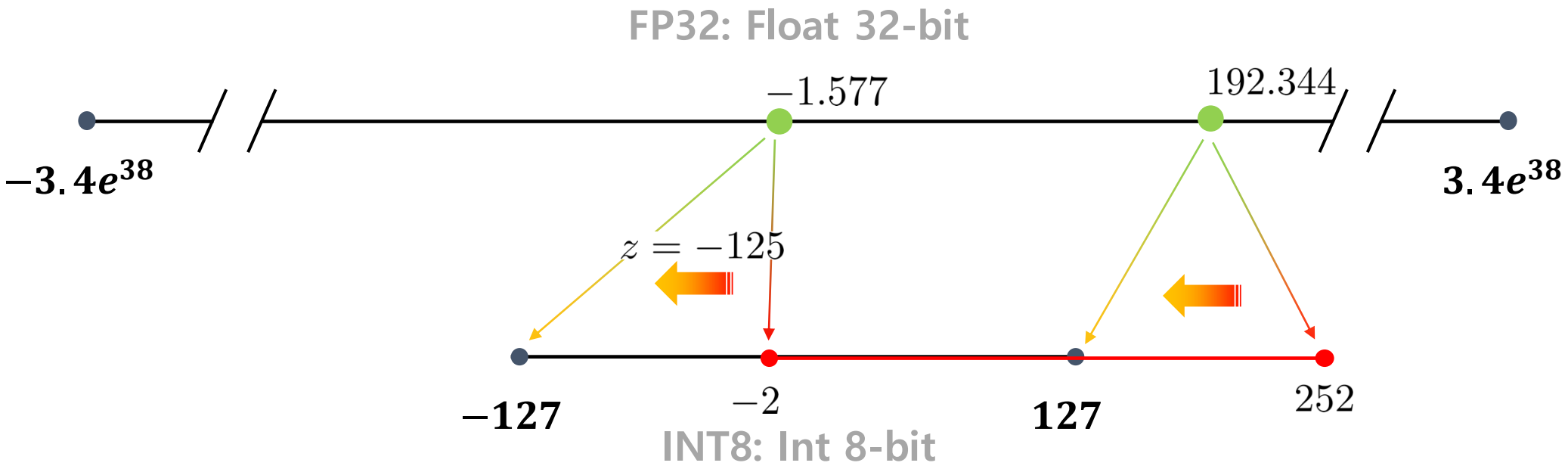
Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

Zero-point integer

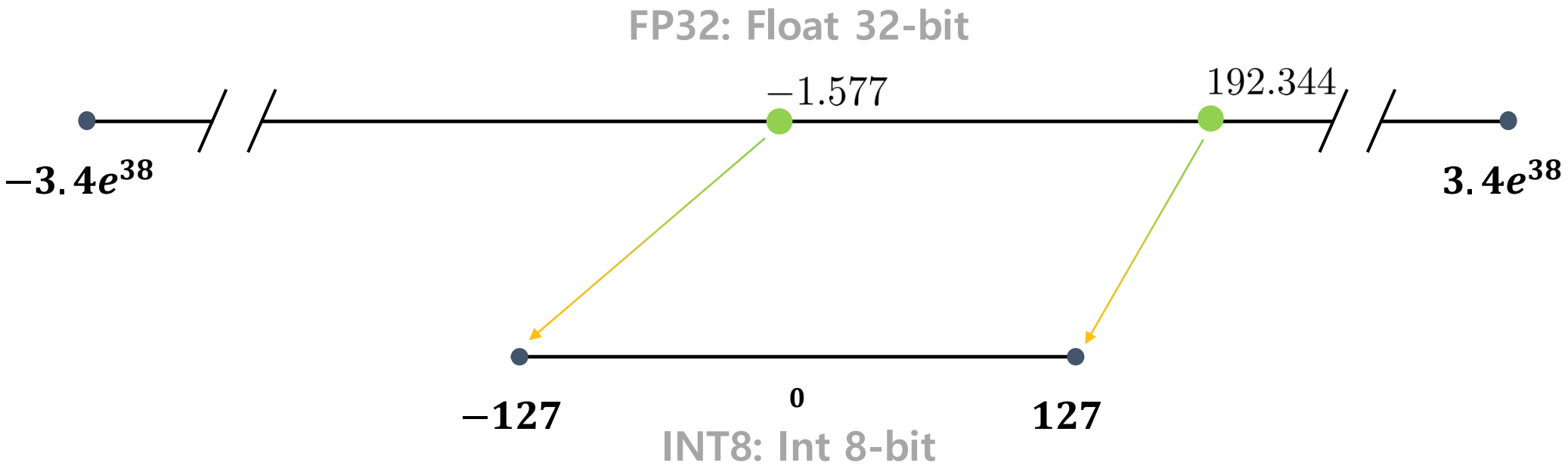


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

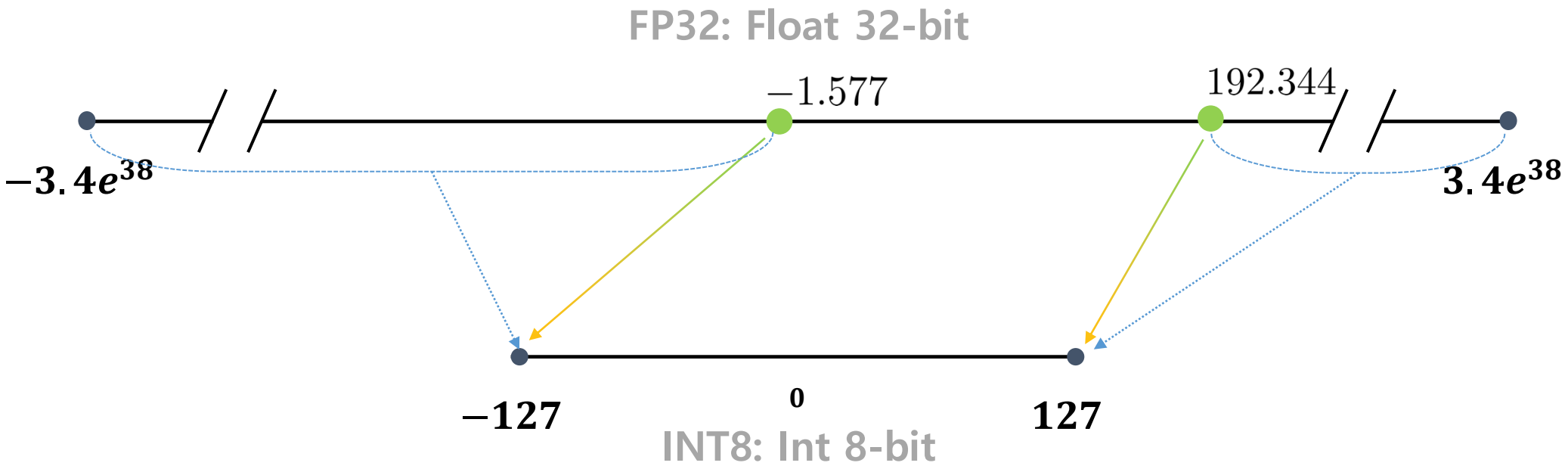


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

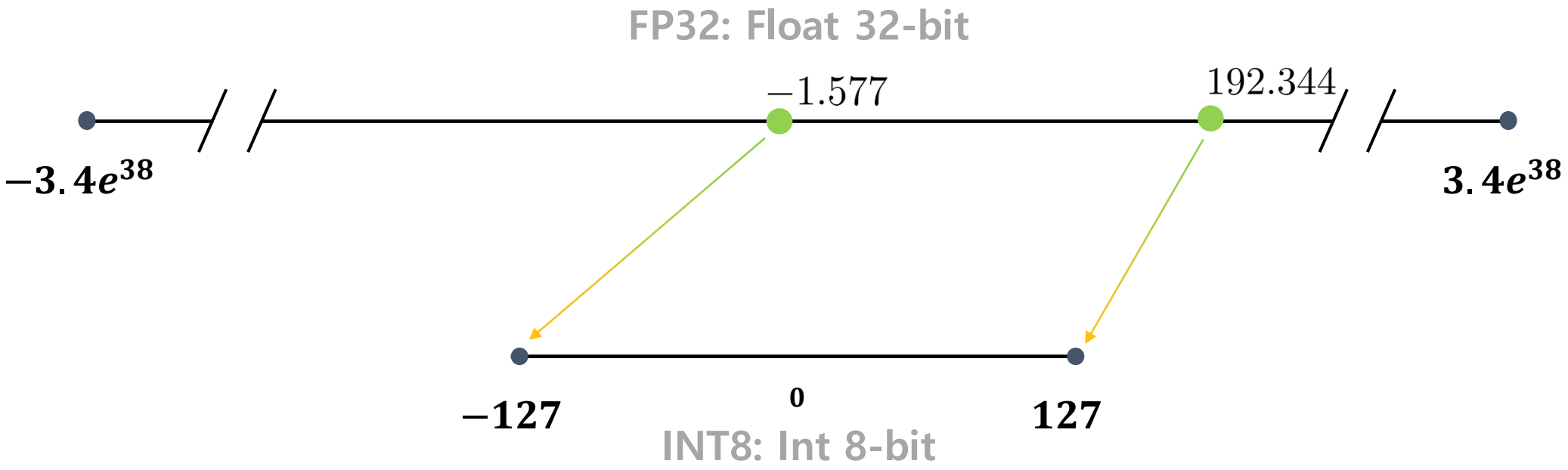


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

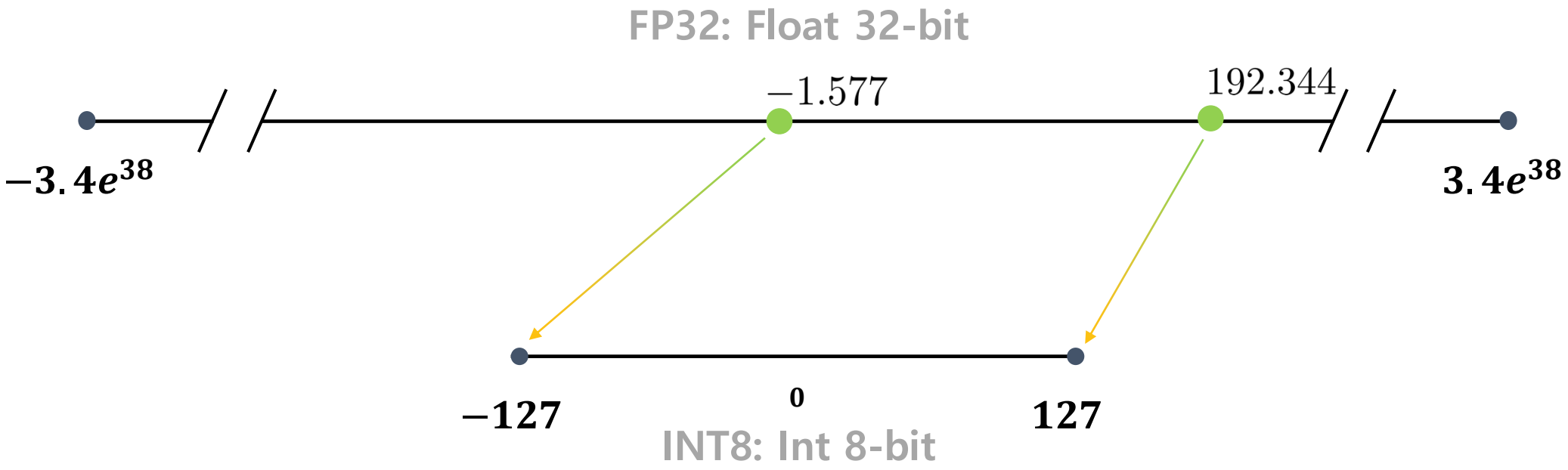


Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

Activation이나 weight 값이 바뀌면, 결과값도 달라질 텐데 괜찮은 걸까?



Basic of Quantization

❖ Quantization Function

➢ 어떻게 실수를 정수로 변환할까?

🤔 **Activation이나 weight 값이 바뀌면, 결과값도 달라질 텐데 괜찮은 걸까?**

실제 지구를
종이 스케일로 맵핑

Basic of Quantization

❖ Quantization Function

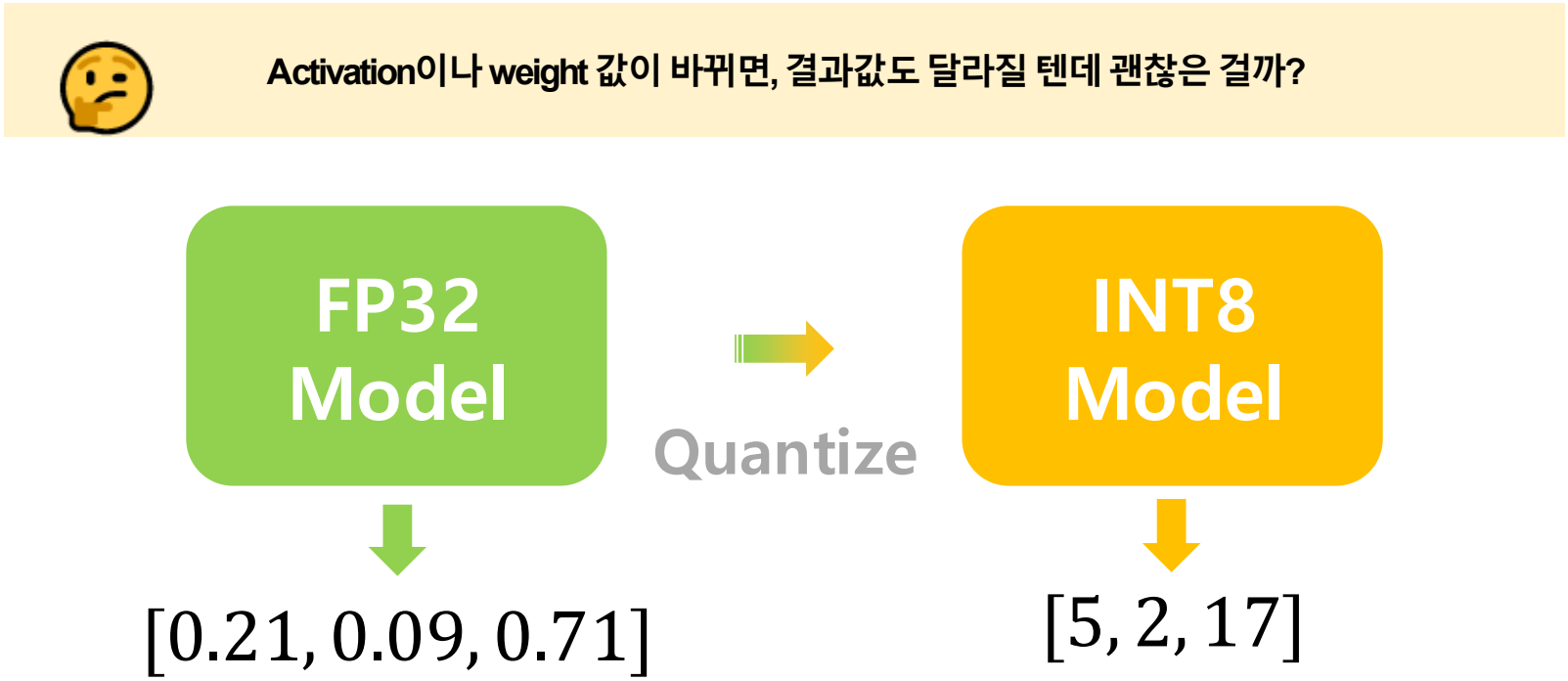
➢ 어떻게 실수를 정수로 변환할까?

Activation이나 weight 값이 바뀌면, 결과값도 달라질 텐데 괜찮은 걸까?

Basic of Quantization

❖ Quantization Function

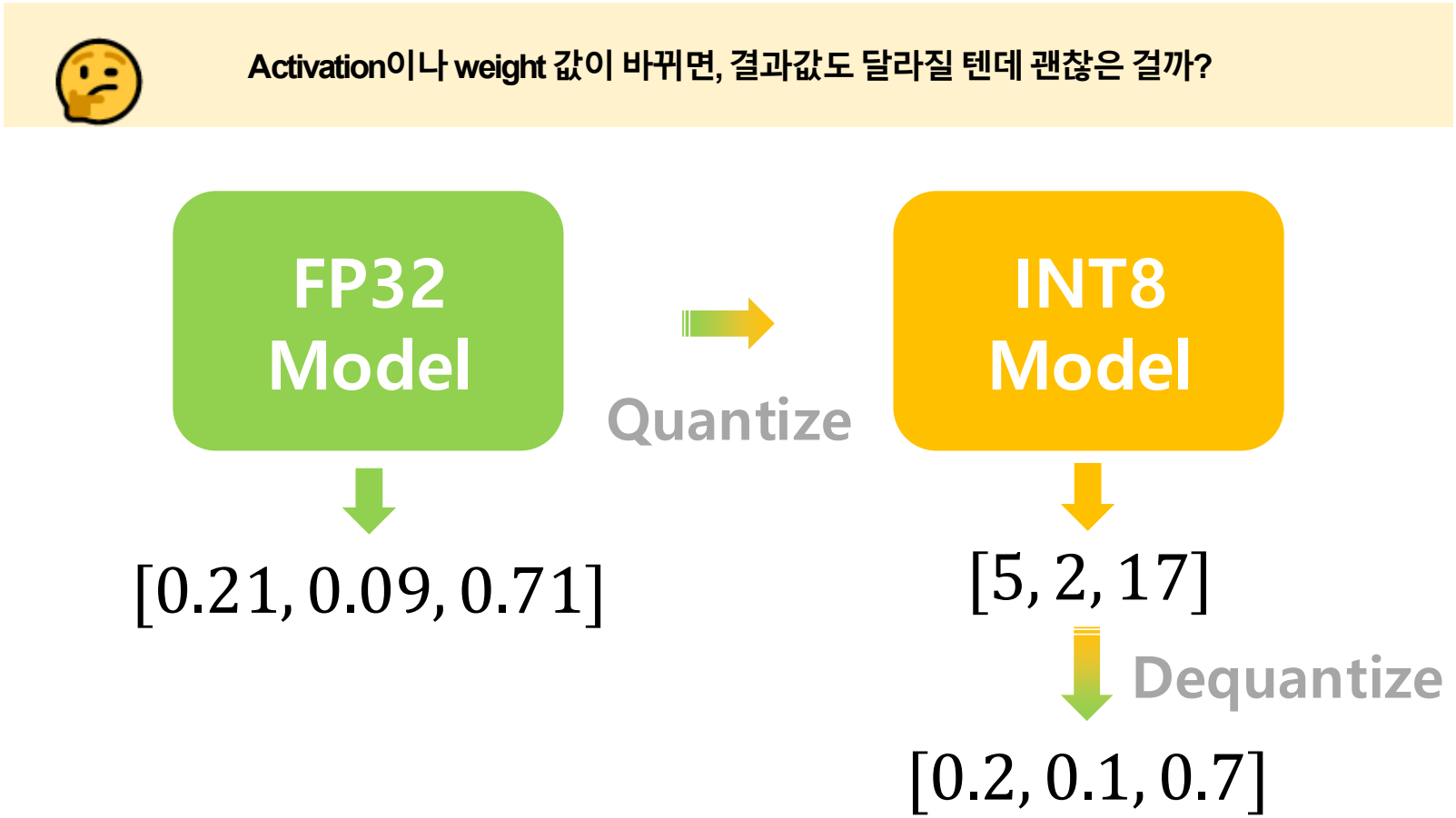
➢ 어떻게 실수를 정수로 변환할까?



Basic of Quantization

❖ Quantization Function

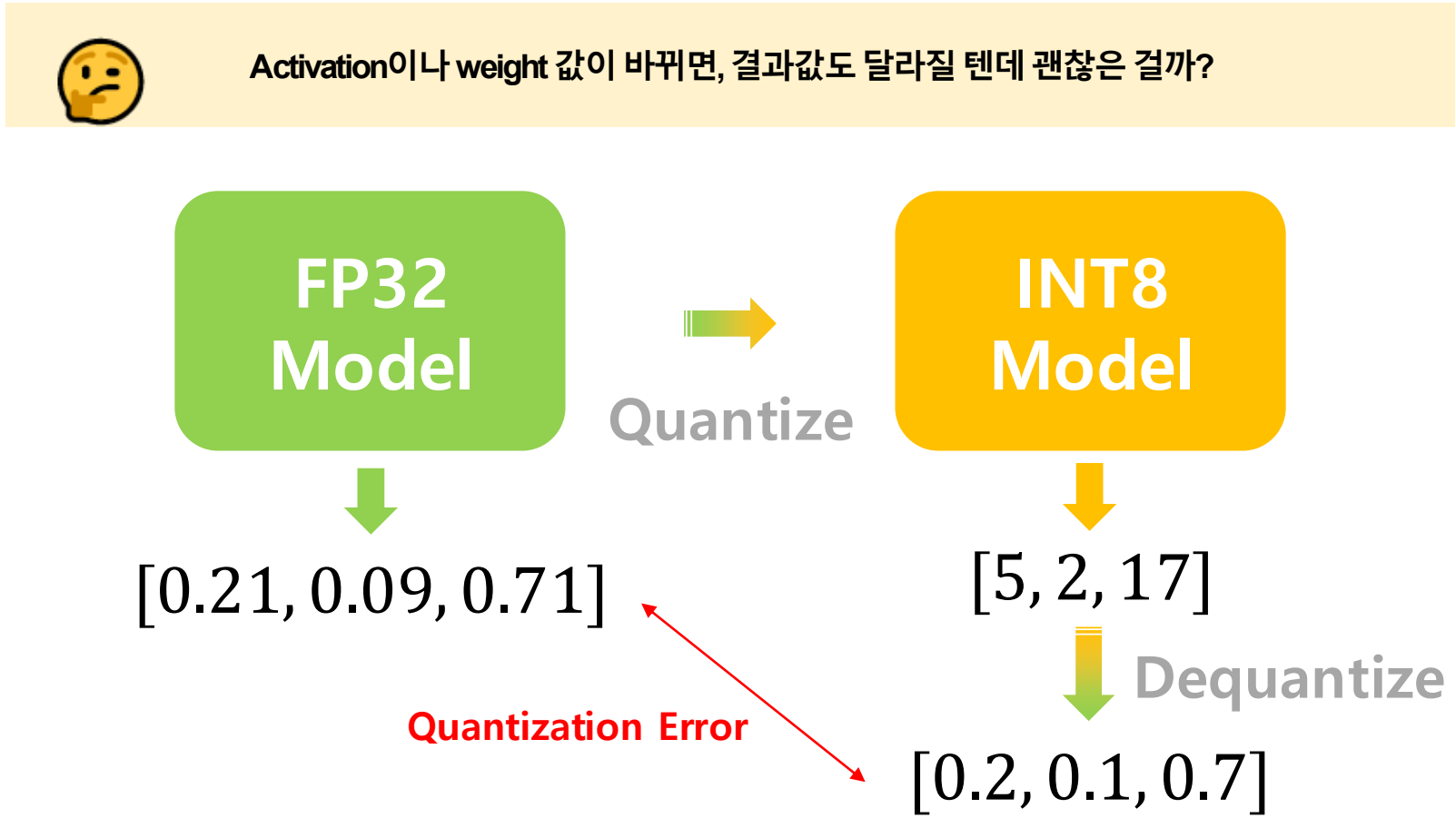
➢ 어떻게 실수를 정수로 변환할까?



Basic of Quantization

❖ Quantization Function

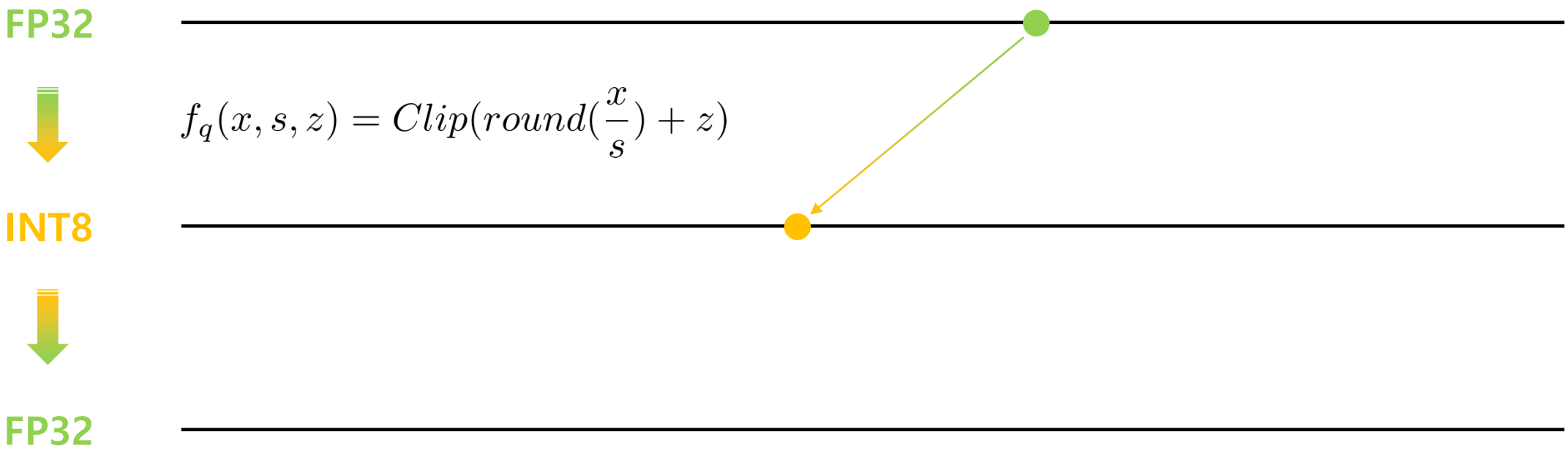
➢ 어떻게 실수를 정수로 변환할까?



Basic of Quantization

❖ Quantization Error

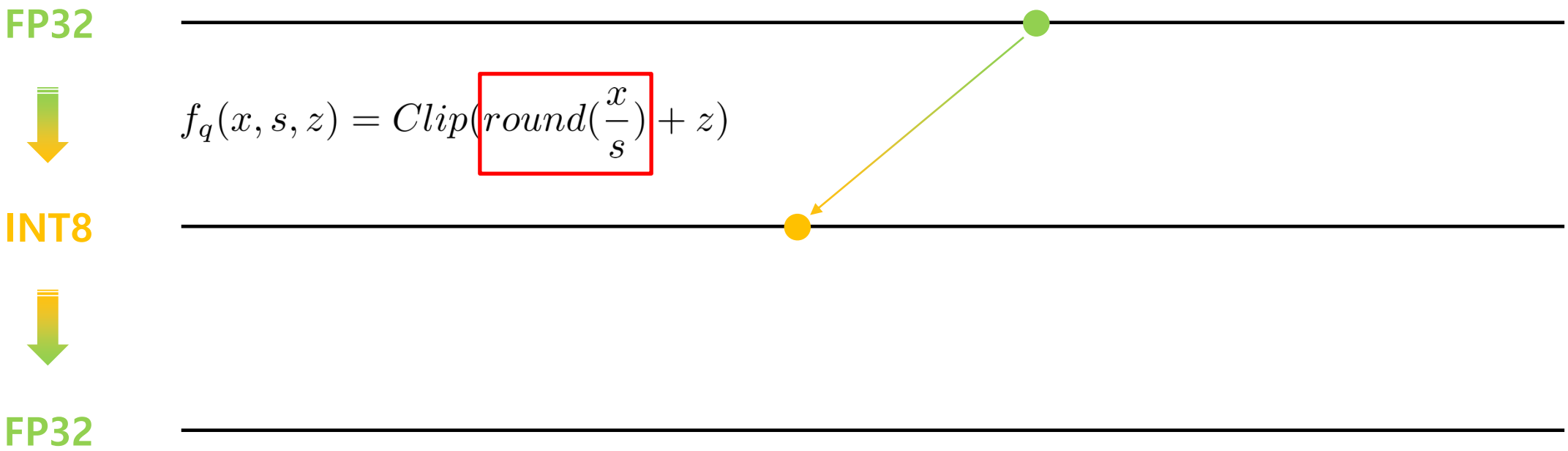
➢ Error를 줄이려면 어떻게 해야 할까?



Basic of Quantization

❖ Quantization Error

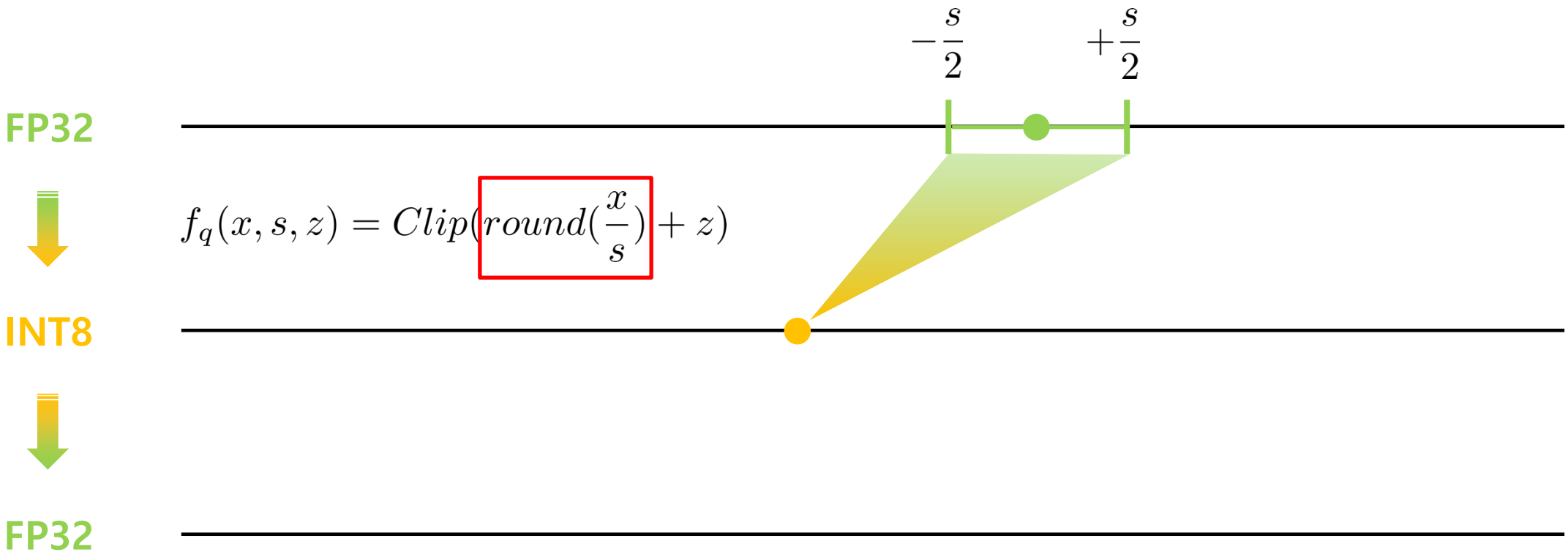
➢ Error를 줄이려면 어떻게 해야 할까?



Basic of Quantization

❖ Quantization Error

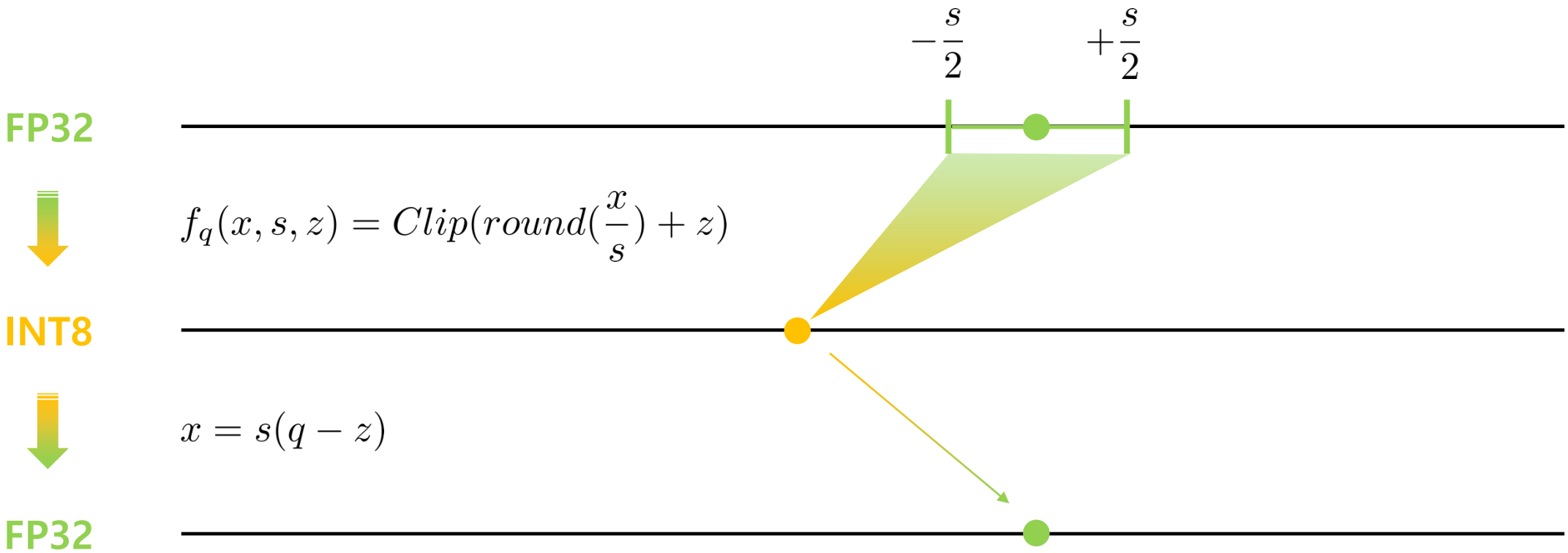
➢ Error를 줄이려면 어떻게 해야 할까?



Basic of Quantization

❖ Quantization Error

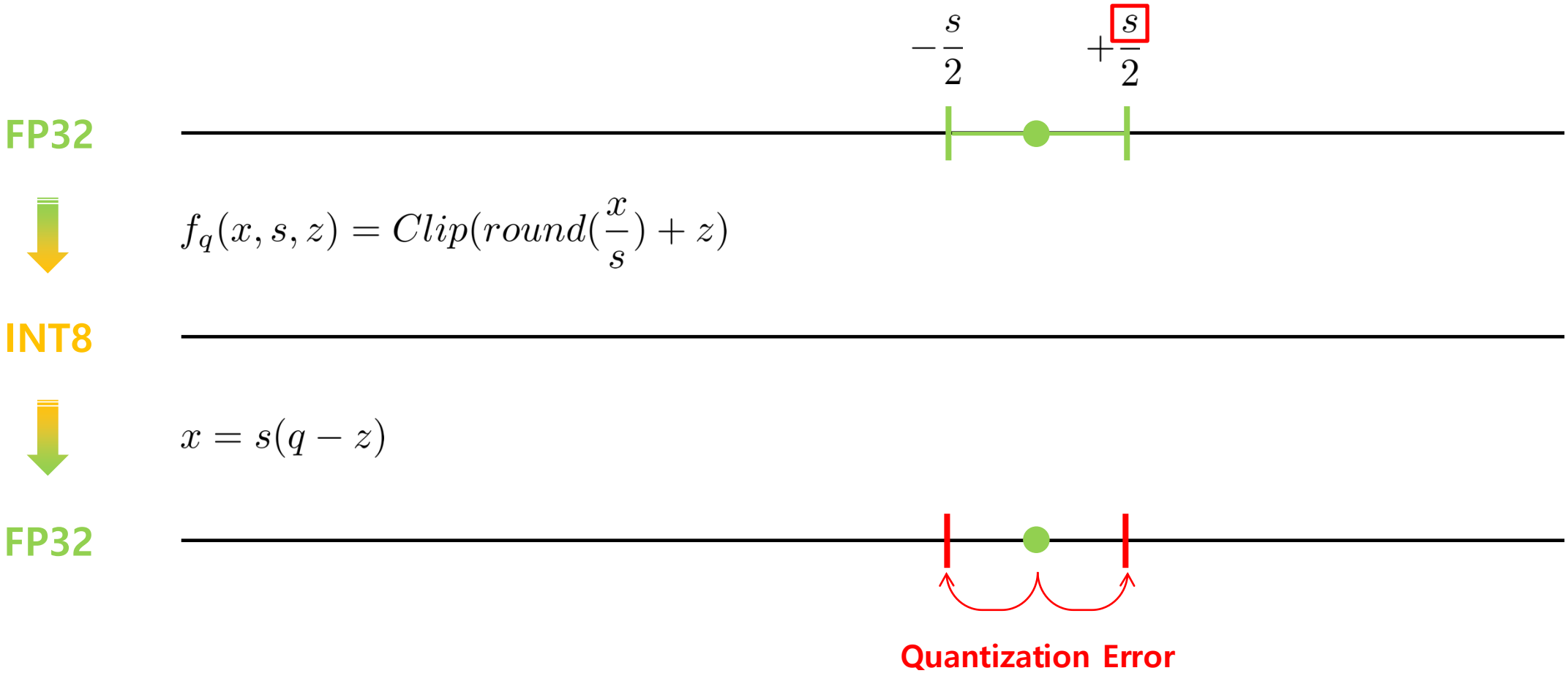
➢ Error를 줄이려면 어떻게 해야 할까?



Basic of Quantization

❖ Quantization Error

➢ Error를 줄이려면 어떻게 해야 할까? → ***s*를 줄여야 한다.**



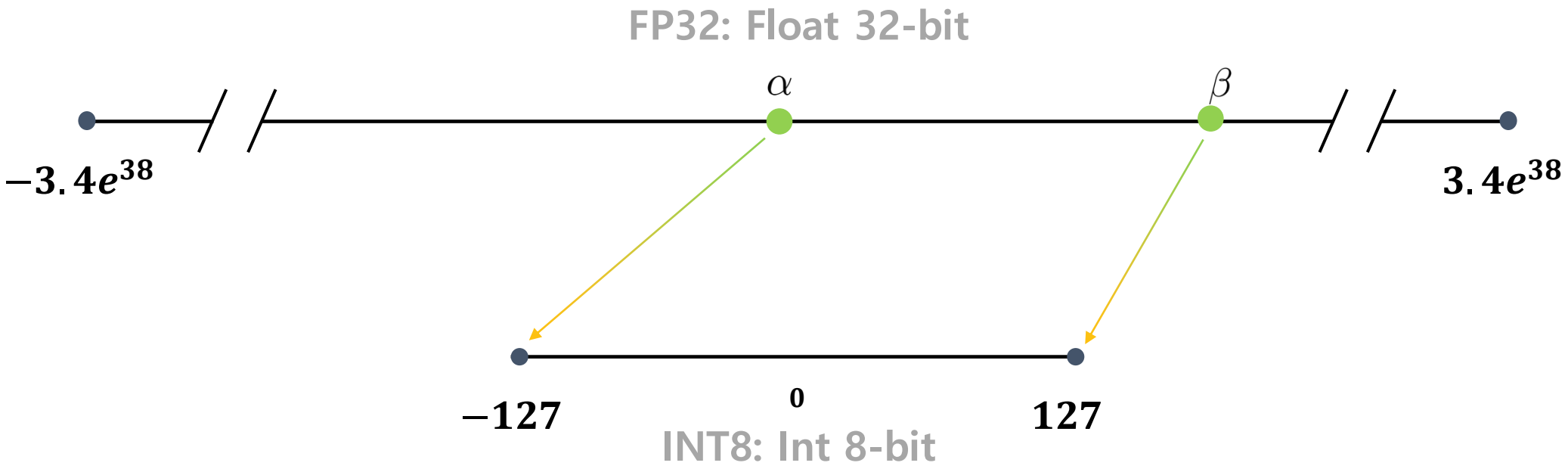
Basic of Quantization

❖ Quantization Error

➢ Error를 줄이려면 어떻게 해야 할까? → ***s*를 줄여야 한다.**

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

$$s = \frac{\beta - \alpha}{254}$$



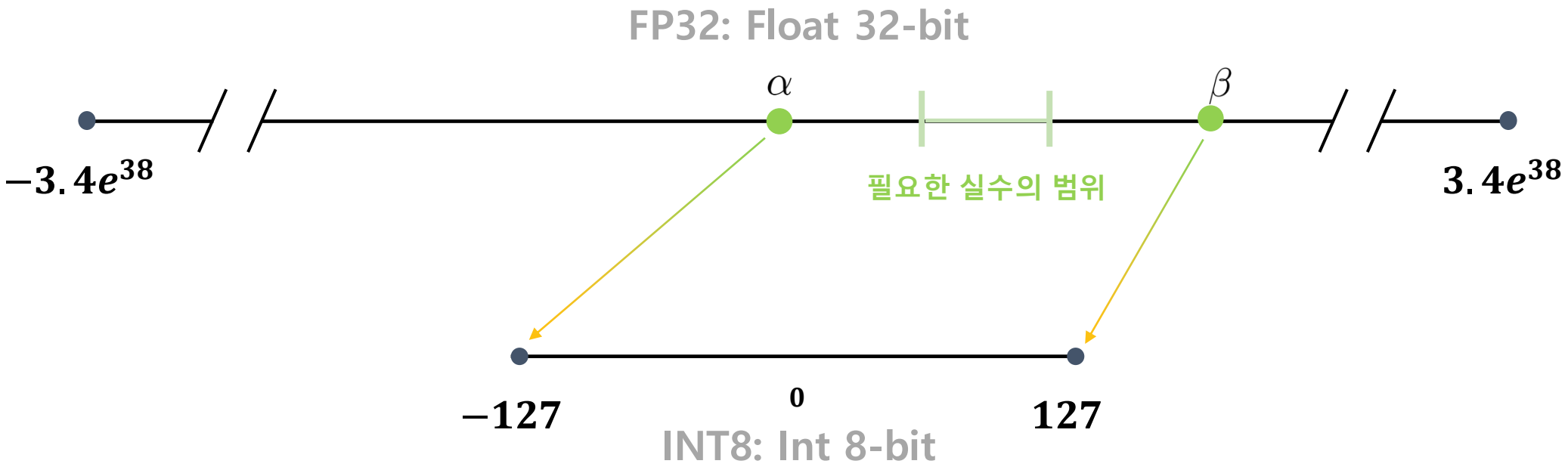
Basic of Quantization

❖ Quantization Error

➢ Error를 줄이려면 어떻게 해야 할까? → ***s*를 줄여야 한다.**

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

$$s = \frac{\beta - \alpha}{254}$$



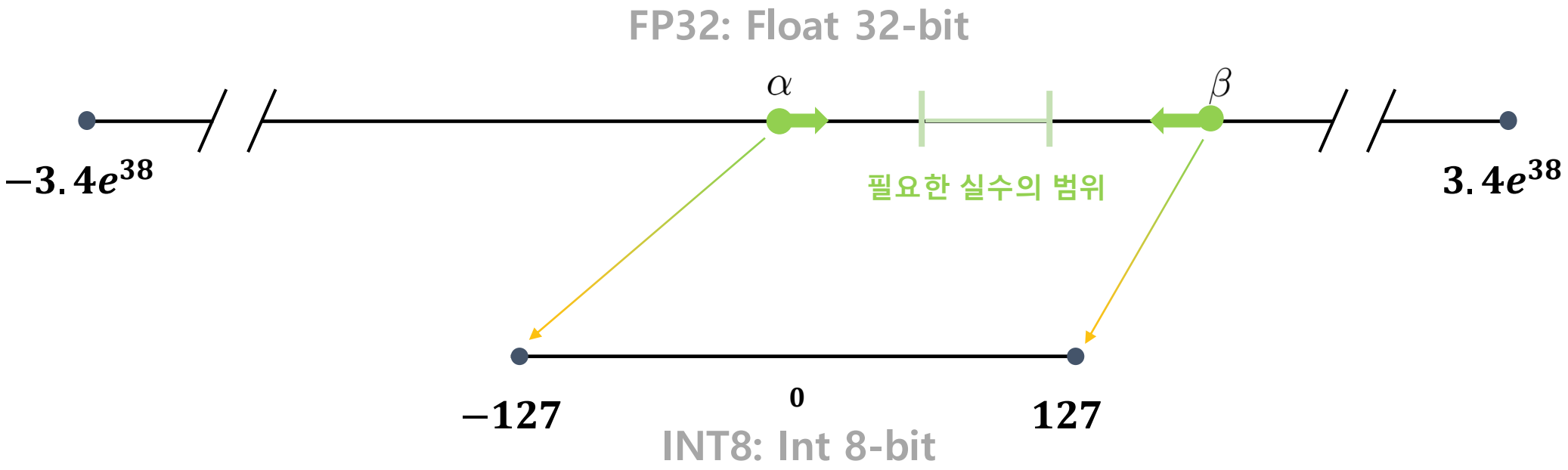
Basic of Quantization

❖ Quantization Error

➢ Error를 줄이려면 어떻게 해야 할까? → ***s*를 줄여야 한다.**

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

$$s = \frac{\beta - \alpha}{254}$$



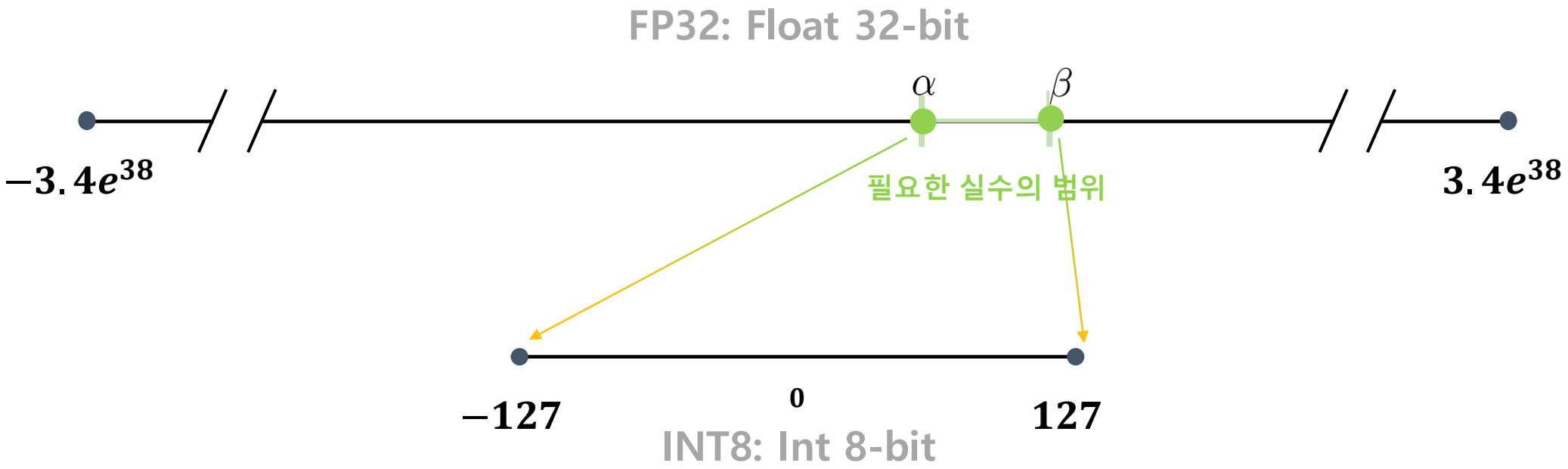
Basic of Quantization

❖ Quantization Error

➢ Error를 줄이려면 어떻게 해야 할까? → ***s*를 줄여야 한다.**

$$f_q(x, s, z) = Clip(round(\frac{x}{s}) + z)$$

$$s = \frac{\beta - \alpha}{254}$$

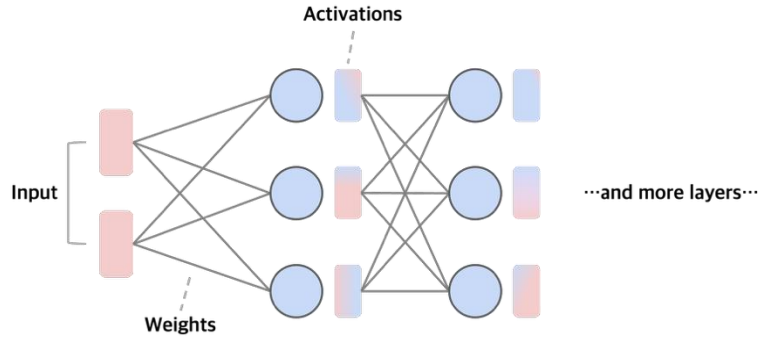


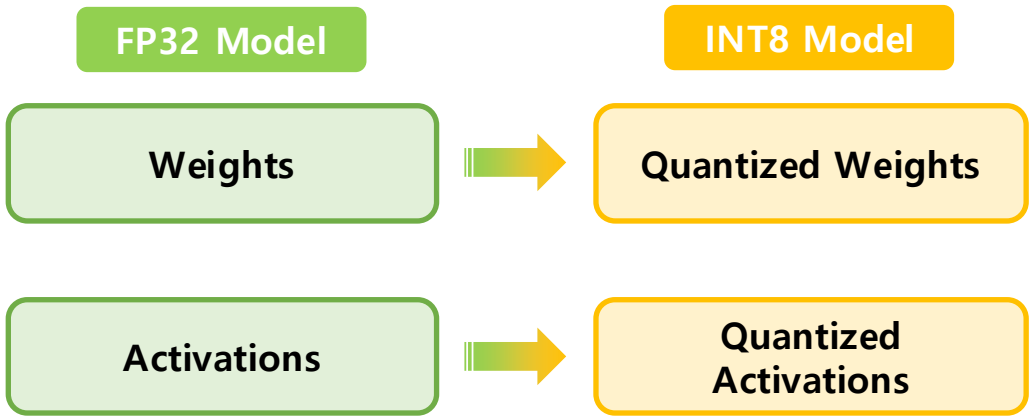
Basic of Quantization

❖ Quantization Error

- ➢ Error를 줄이려면 어떻게 해야 할까? → s**를 줄여야 한다.**
- ➢ **필요한 실수의 범위를 정확히 파악해야 한다.**

$$s = \frac{\beta - \alpha}{254}$$

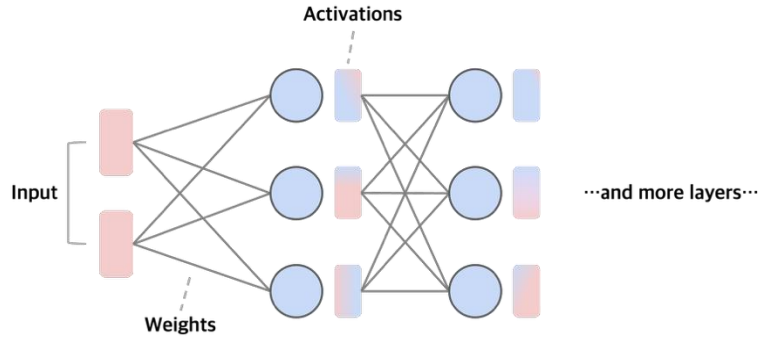




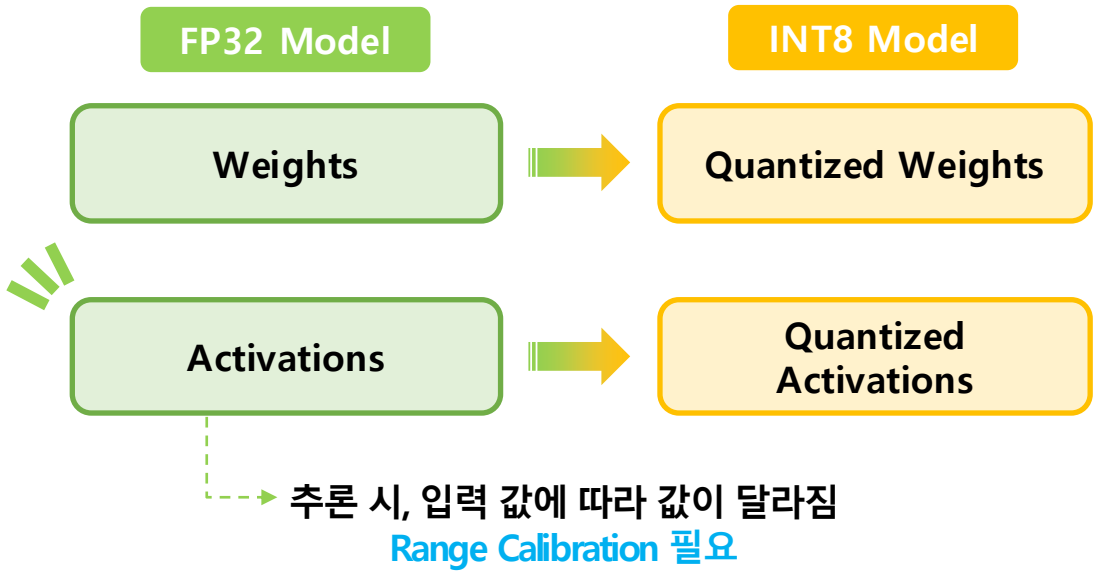
Basic of Quantization

❖ Quantization Error

- ➢ Error를 줄이려면 어떻게 해야 할까? → s**를 줄여야 한다.**
- ➢ **필요한 수의 범위를 정확히 파악해야 한다.**



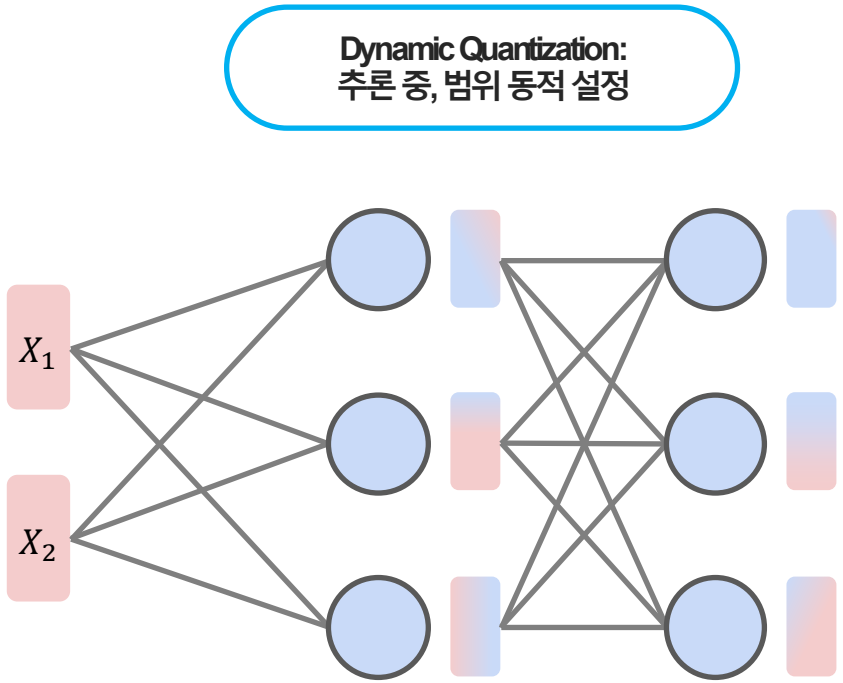
$$s = \frac{\beta - \alpha}{254}$$



Range Calibration

❖ Post Training Quantization (PTQ)

- ➢ Quantization after training
  - Dynamic Quantization
  - Static Quantization

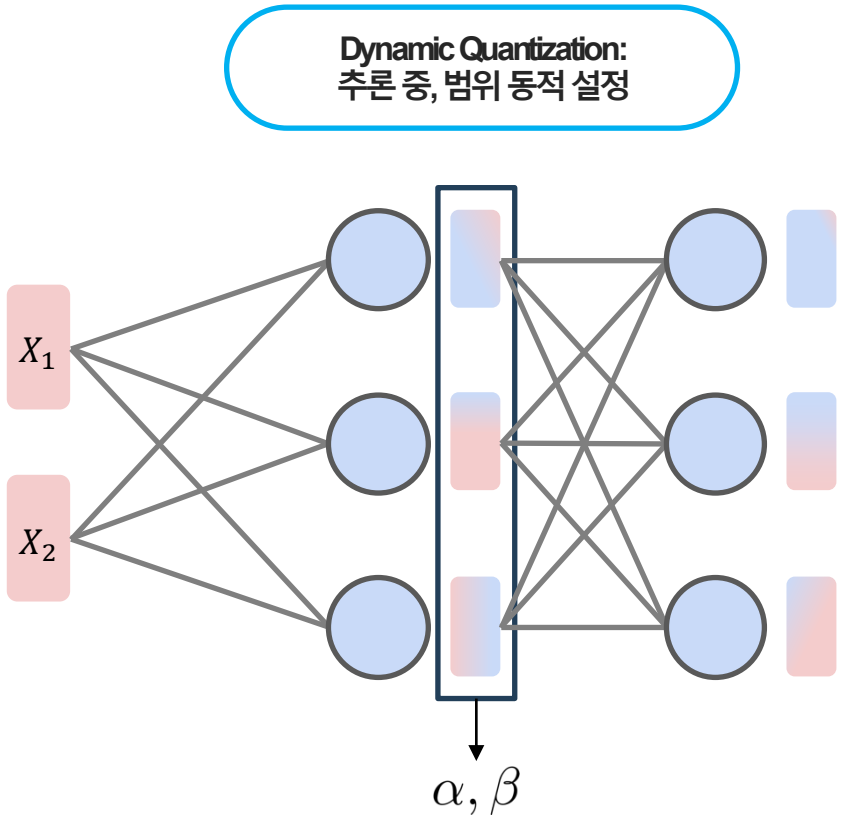


Range Calibration

❖ Post Training Quantization (PTQ)

➢ Quantization after training

- Dynamic Quantization
- Static Quantization

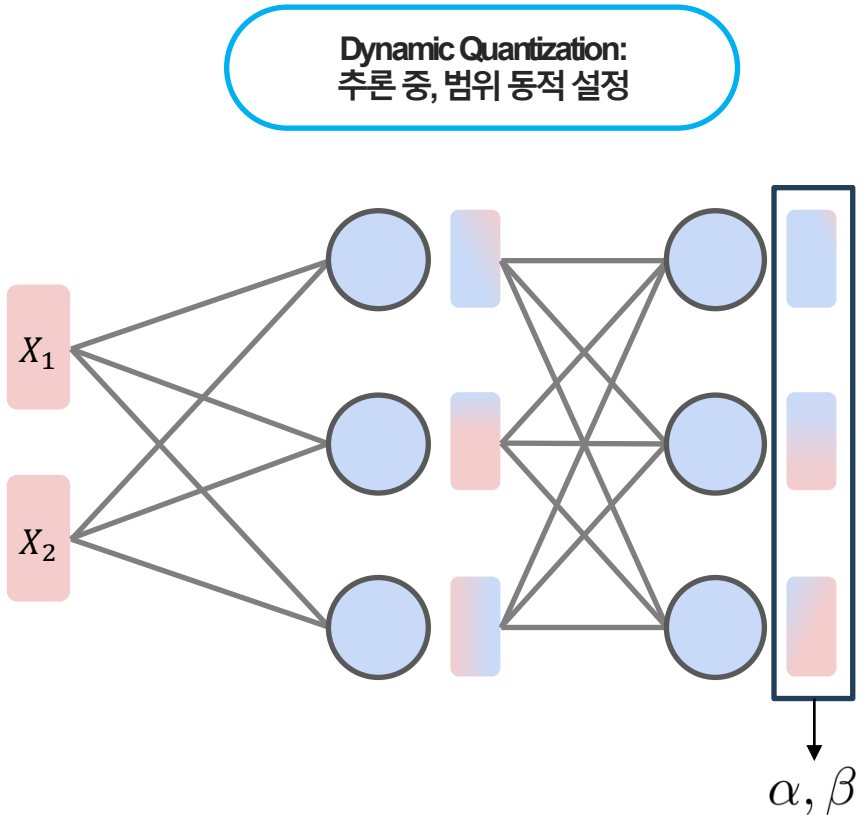


Range Calibration

❖ Post Training Quantization (PTQ)

➢ Quantization after training

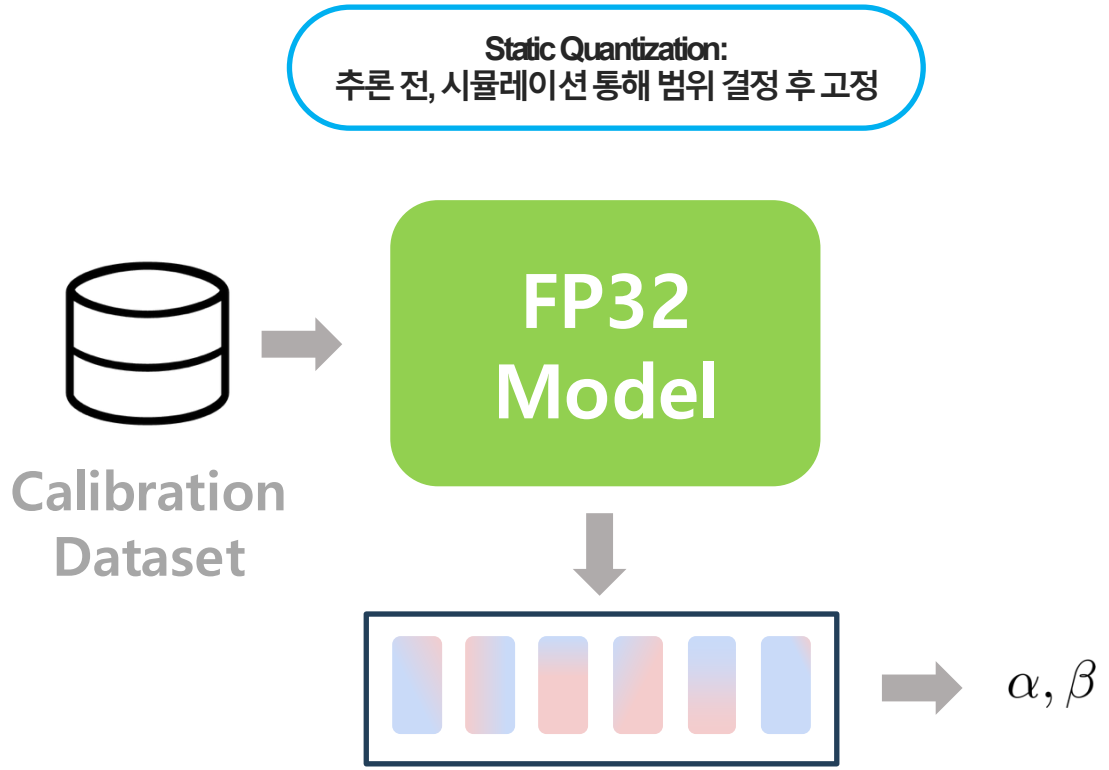
- Dynamic Quantization
- Static Quantization



Range Calibration

❖ Post Training Quantization (PTQ)

- ➢ Quantization after training
  - Dynamic Quantization
  - Static Quantization



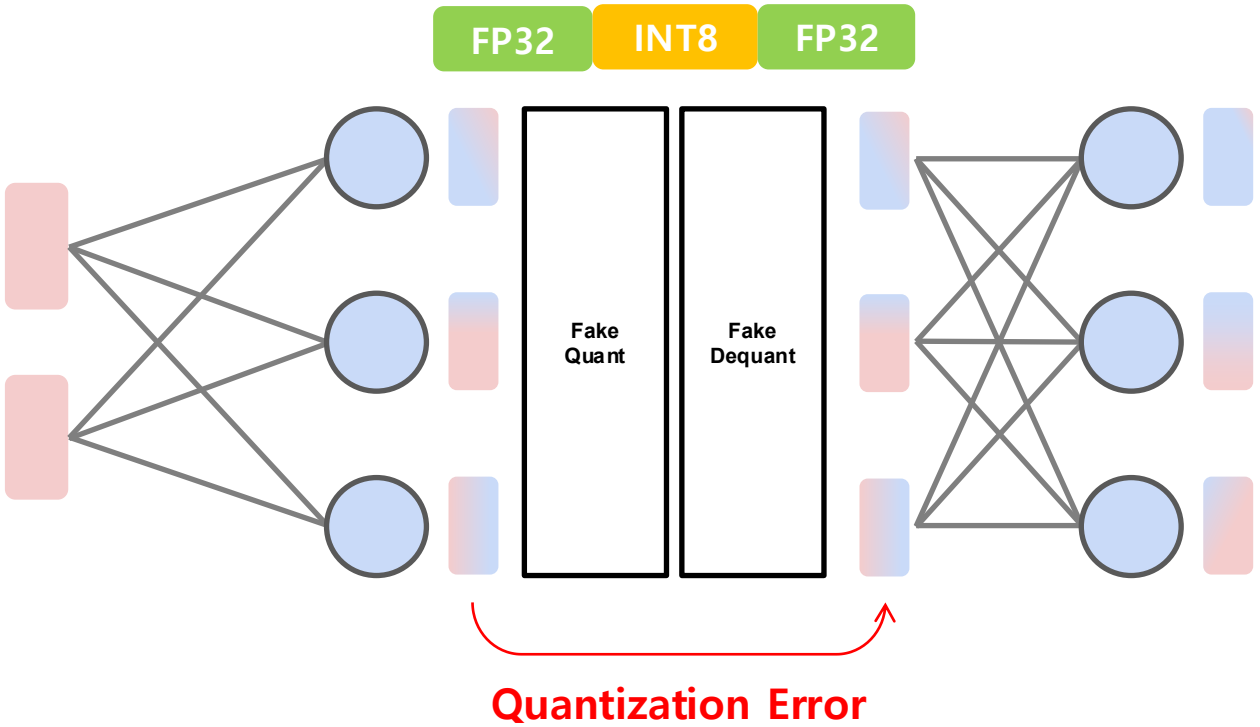
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 학습 수행

① PTQ와 달리 적합한 α, β를 weight과 함께 직접 학습

② 모델이 quantization error에 적응



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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).

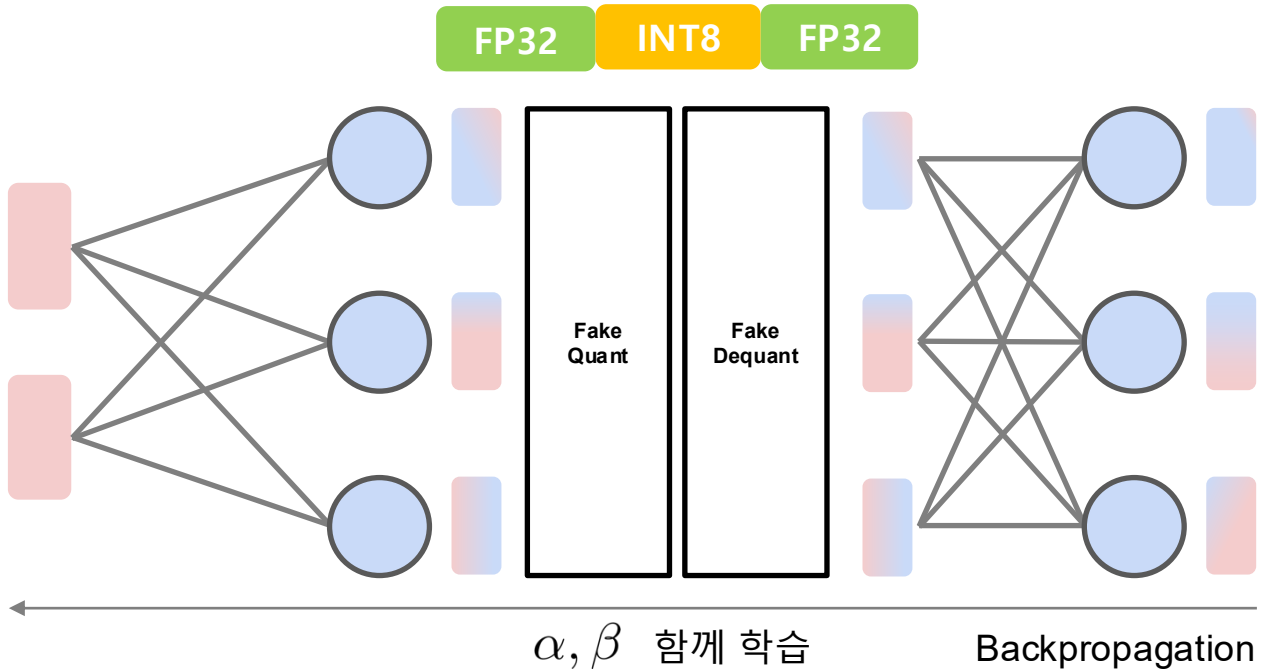
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 학습 수행

① PTQ와 달리 적합한 α, β를 **weight**과 함께 직접 학습

② 모델이 quantization error에 적응



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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).

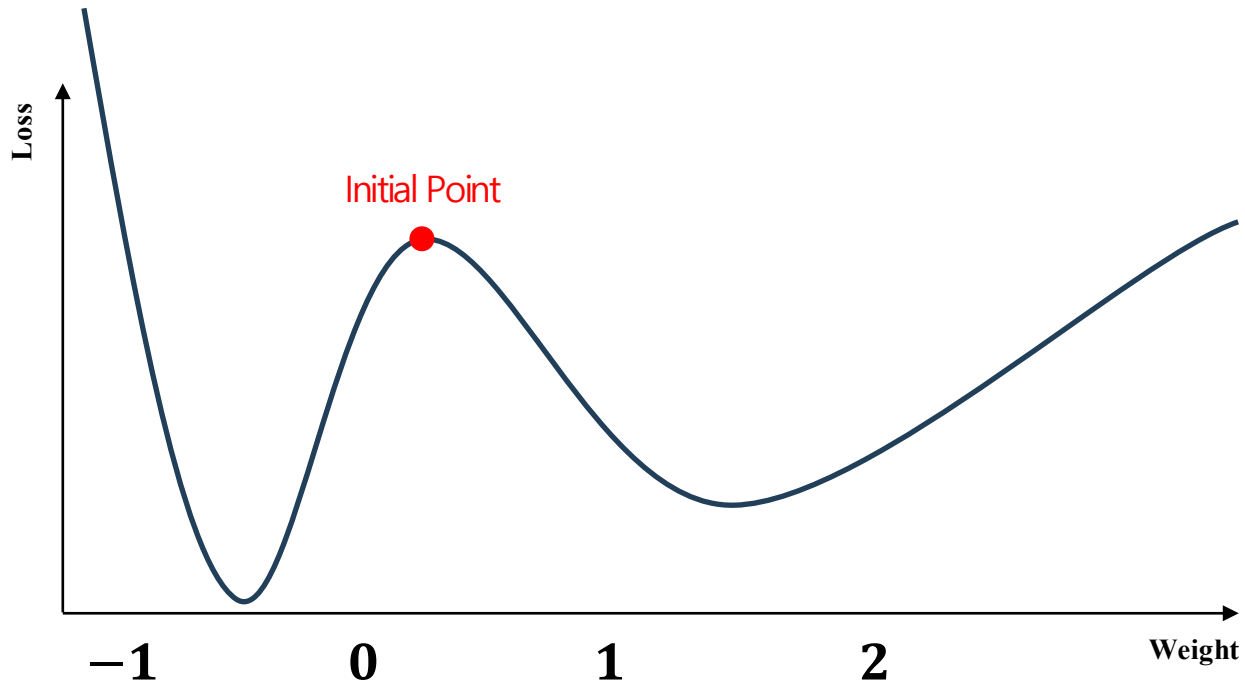
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 미세조정 수행

① PTQ와 달리 적합한 α, β를 **weight**과 함께 직접 학습

② 모델이 quantization error에 적응



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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).

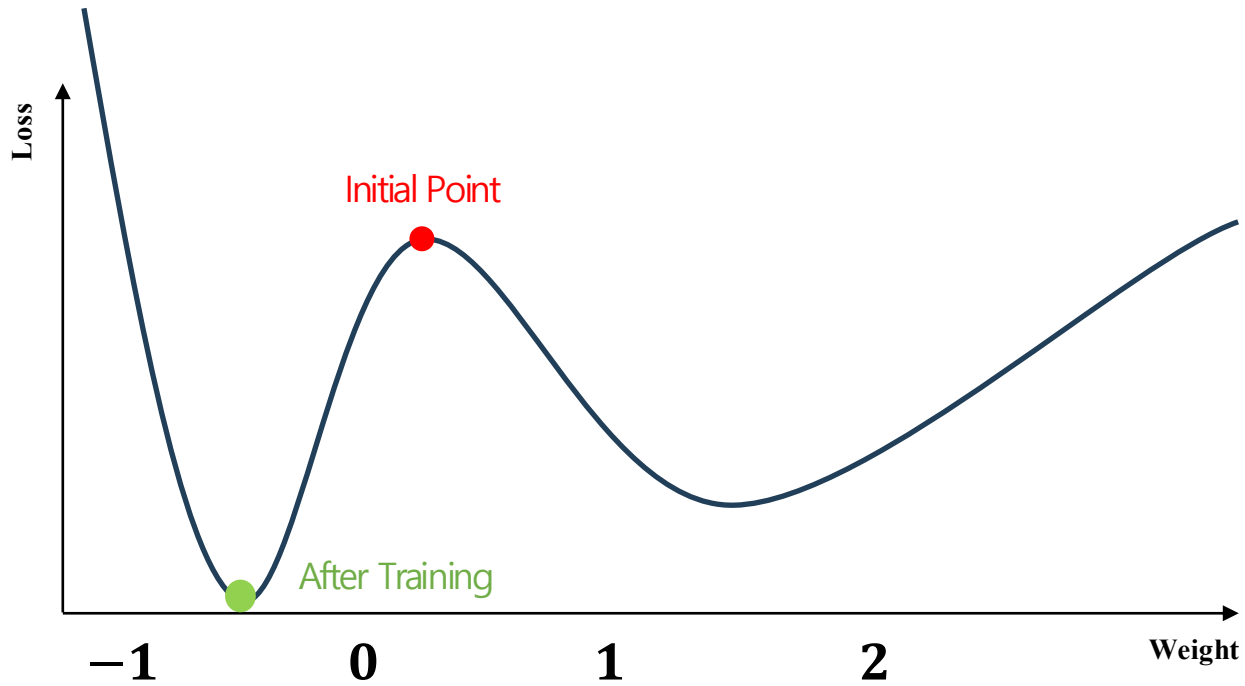
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 미세조정 수행

① PTQ와 달리 적합한 α, β를 **weight**과 함께 직접 학습

② 모델이 quantization error에 적응



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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).

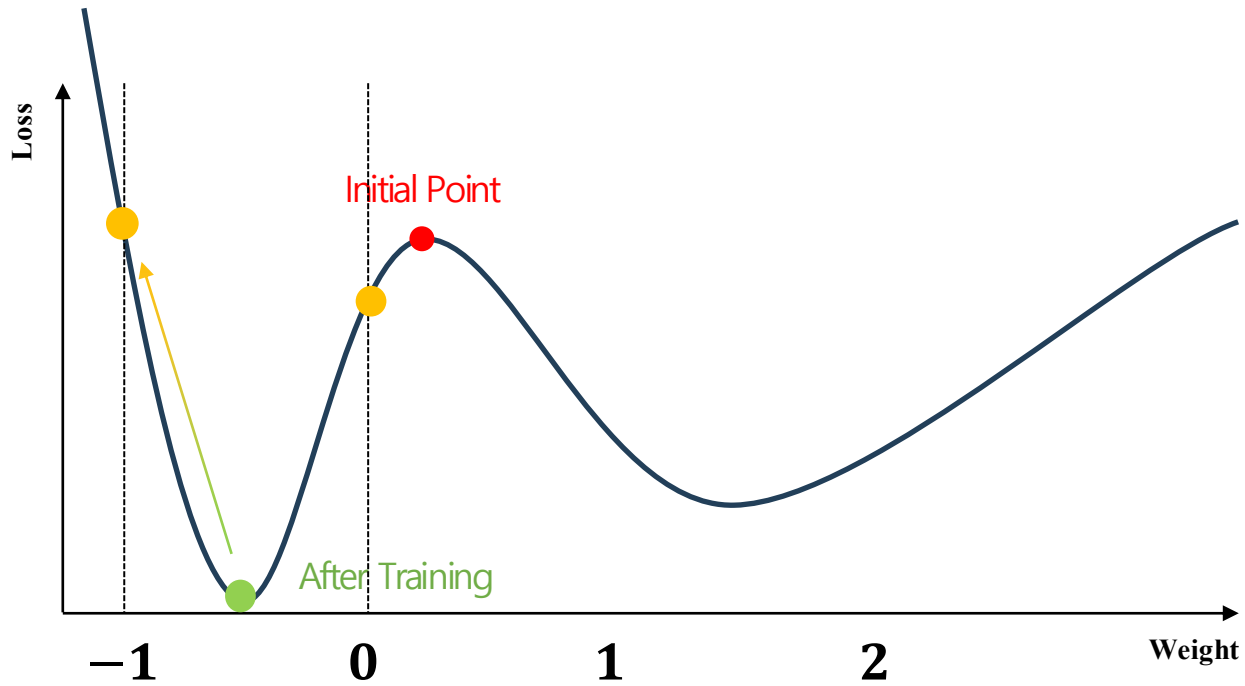
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 미세조정 수행

① PTQ와 달리 적합한 α, β를 **weight**과 함께 직접 학습

② 모델이 quantization error에 적응



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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).

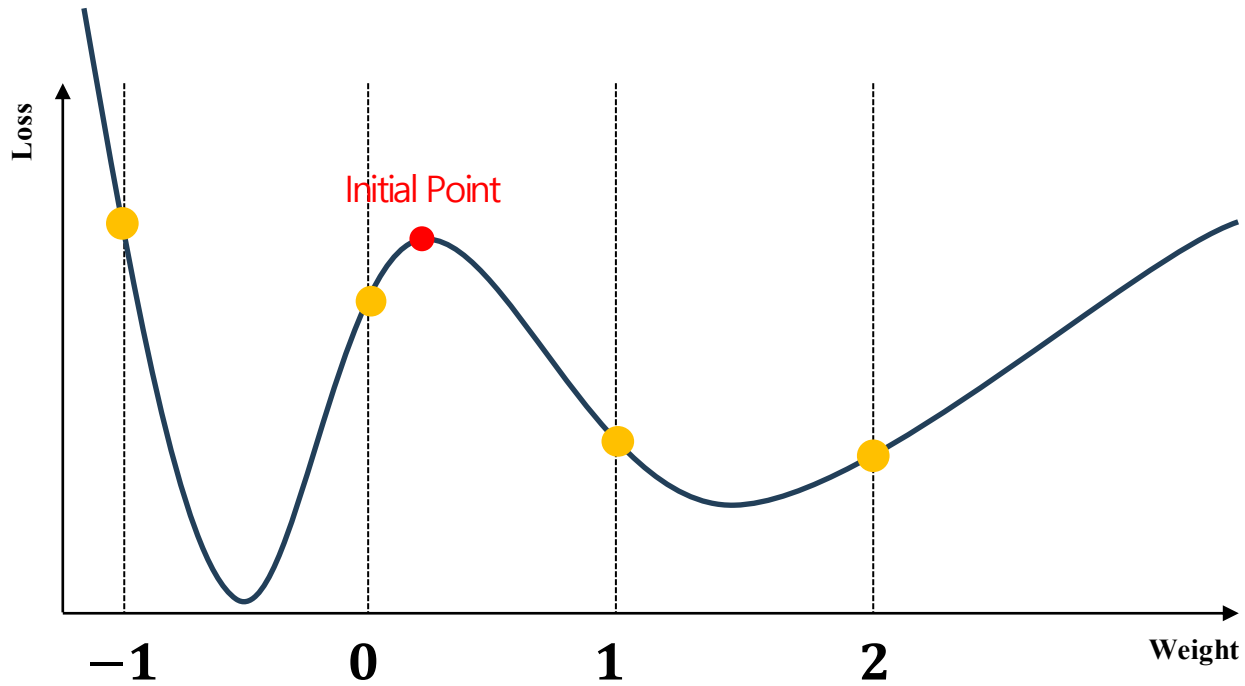
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 미세조정 수행

① PTQ와 달리 적합한 α, β를 **weight**과 함께 직접 학습

② 모델이 quantization error에 적응



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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).

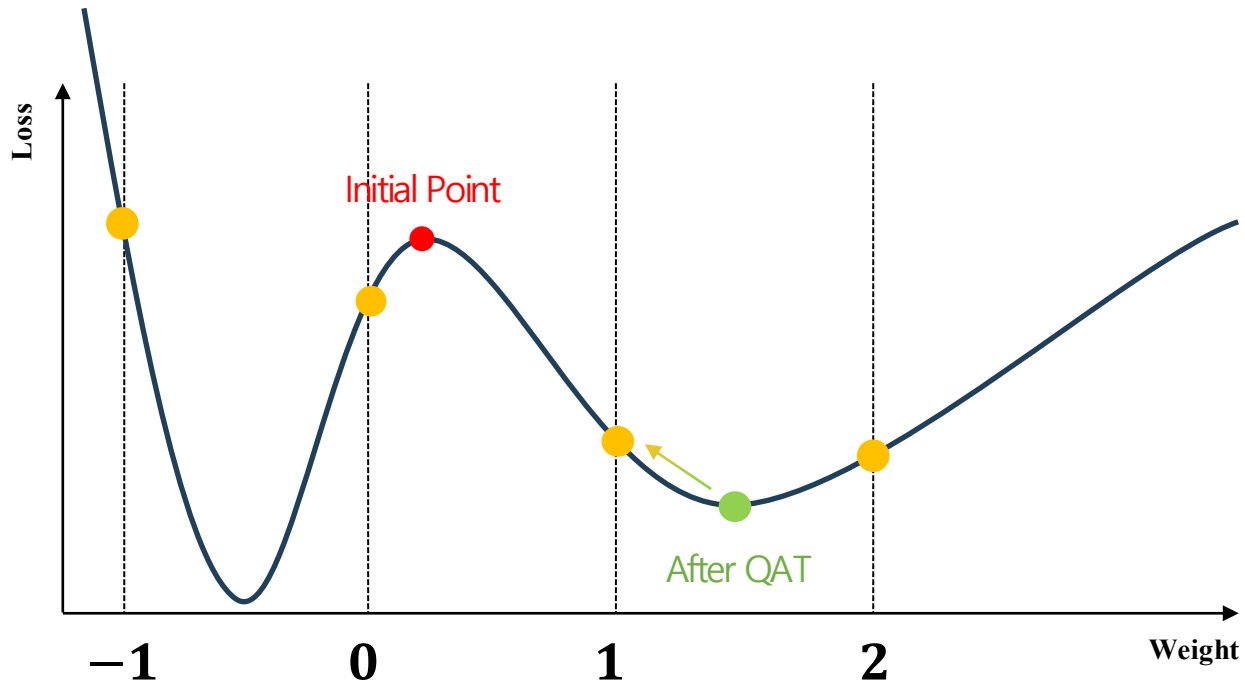
Range Calibration

❖ Quantization Aware Training (QAT)

➢ 추론 전, Fake quantization 도입 및 미세조정 수행

① PTQ와 달리 적합한 α, β를 **weight**과 함께 직접 학습

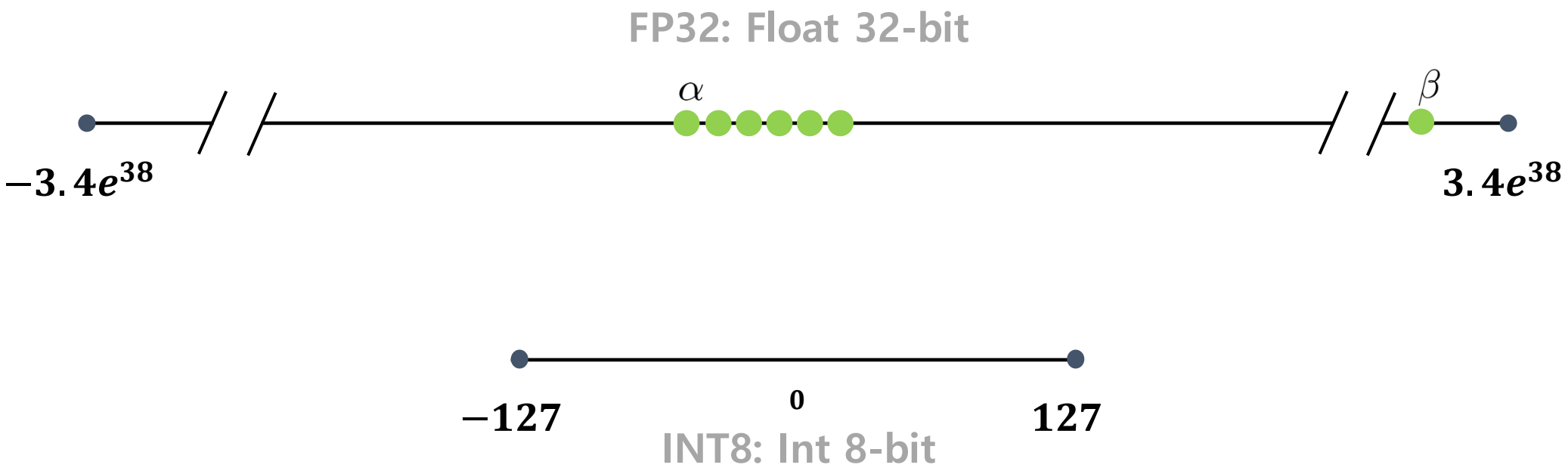
② 모델이 quantization error에 적응



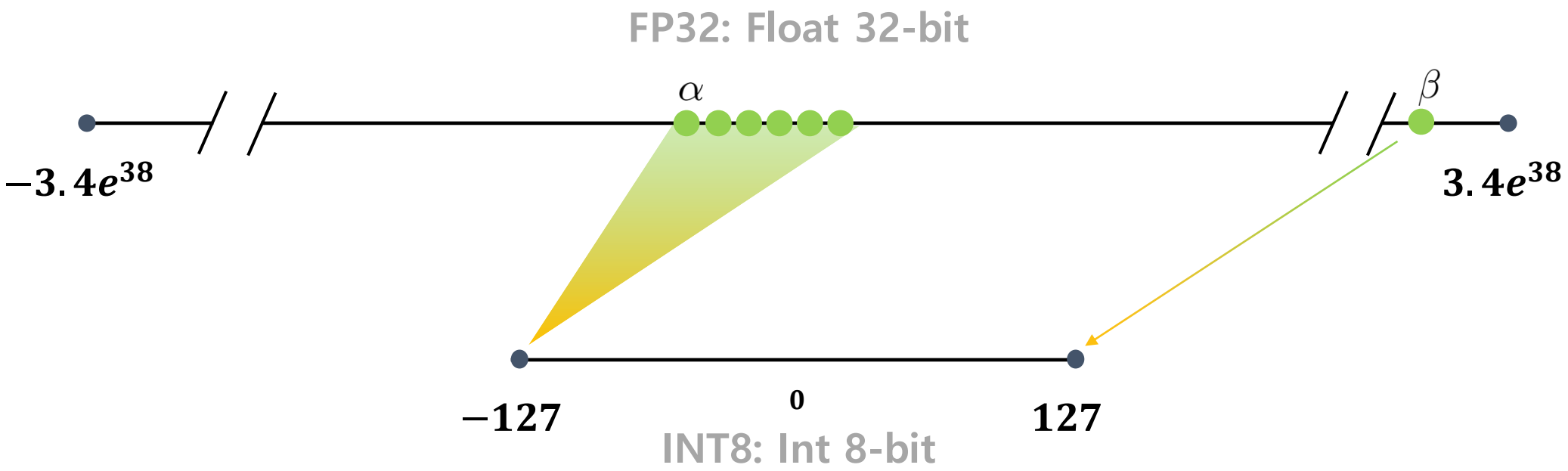
Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2704-2713).

Range Calibration

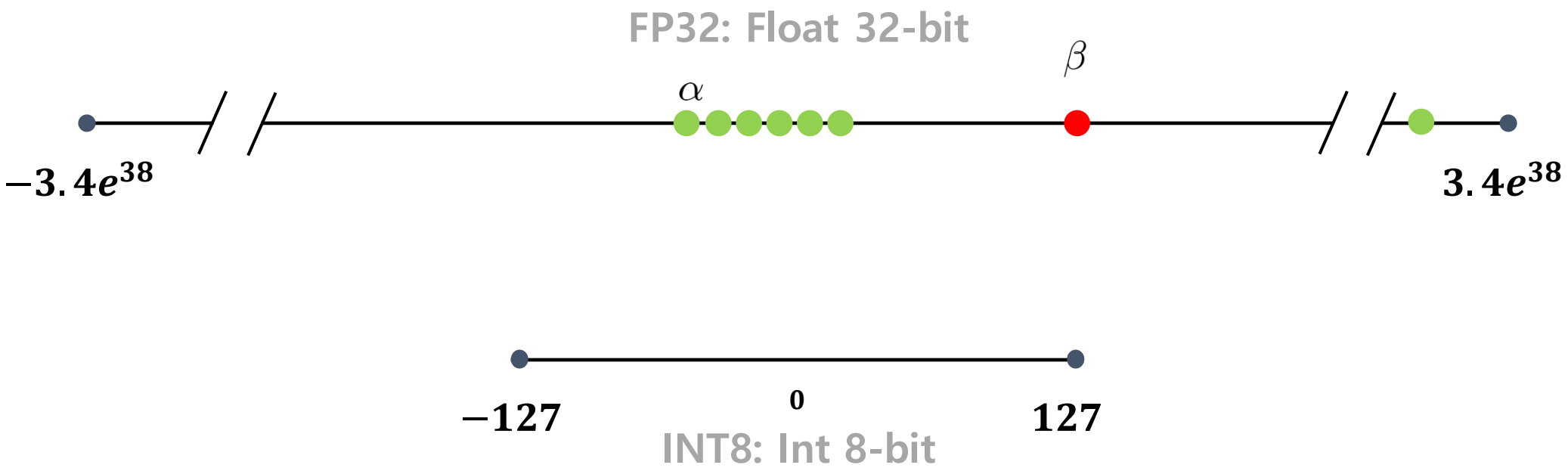
Activation Outliers



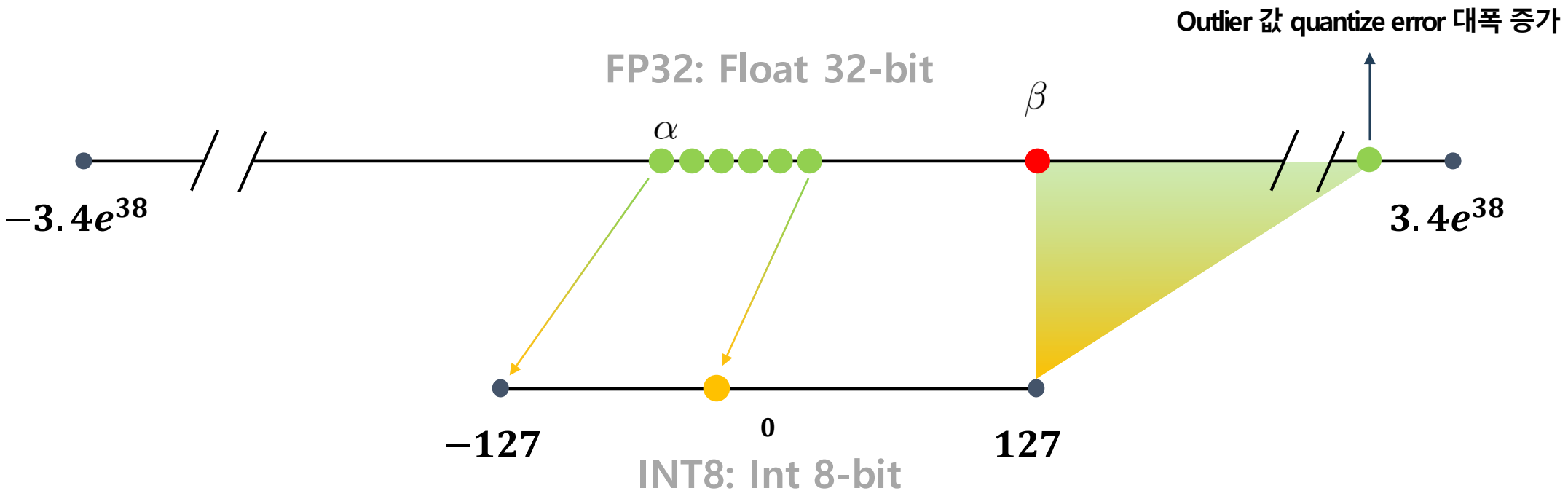
Activation Outliers



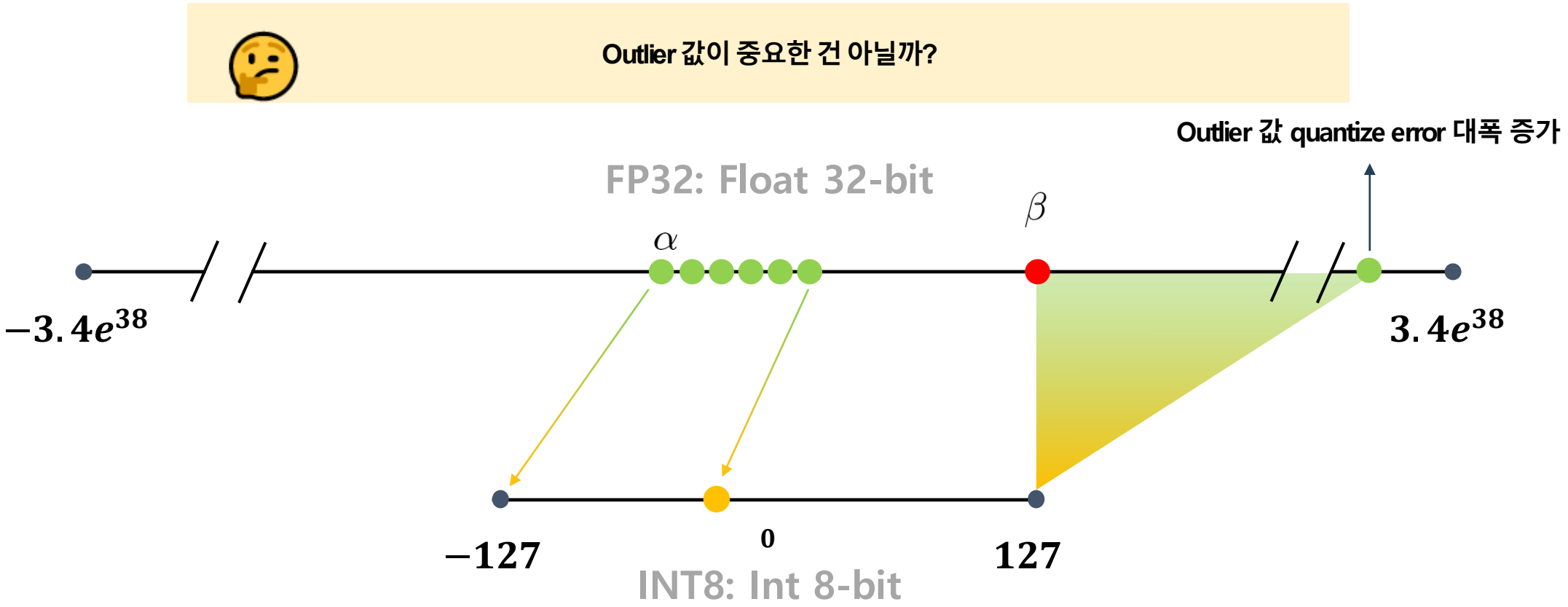
Activation Outliers



Activation Outliers

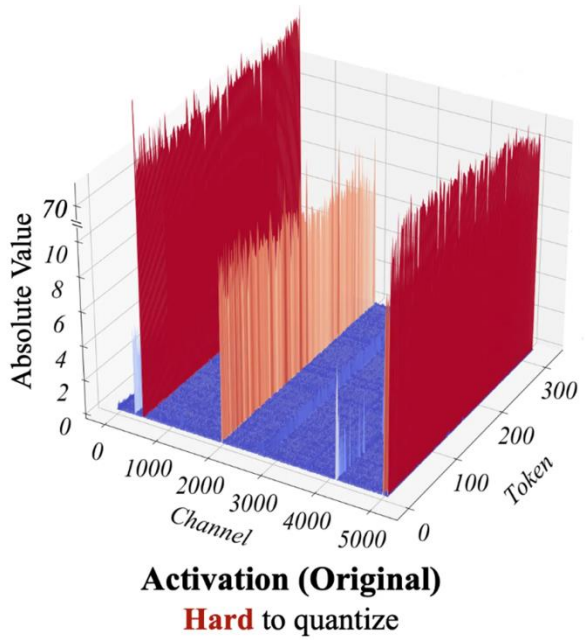


Activation Outliers



Activation Outliers

🤔 **Outlier 값이 중요한 건 아닐까?**

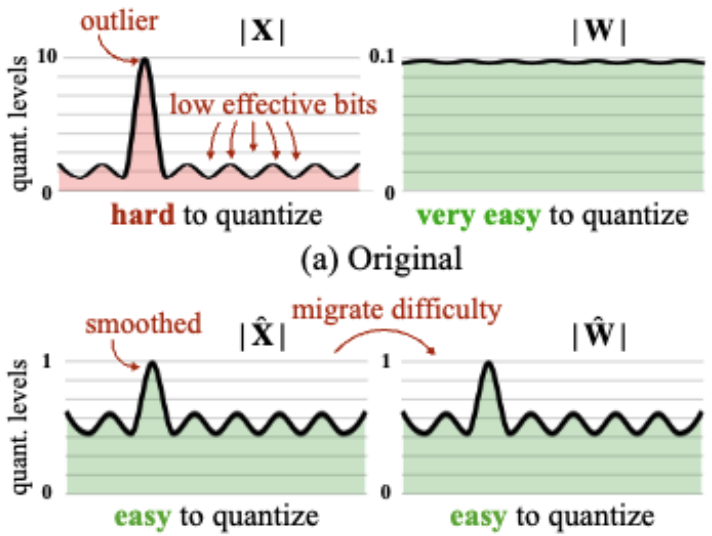


Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023, July). Smoothquant: Accurate and efficient post-training quantization for large language models. In International conference on machine learning (pp. 38087-38099). PMLR.

Activation Outliers

❖ SmoothQuant (PMLR'23)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 변동성이 weight 대비 크다 → Activation Quantization 난이도 ↑
- ➢ 핵심 아이디어: Activation 변동성 일부를 weight로 이전



$$m_j^{(X)} = \max_t |X_{t,j}|, \qquad m_j^{(W)} = \max_k |W_{j,k}|.$$

$$s_j = \frac{(m_j^{(X)})^\alpha}{(m_j^{(W)})^{1-\alpha}}.$$

$$S = \mathrm{diag}(s_1, \ldots, s_{C_i}).$$

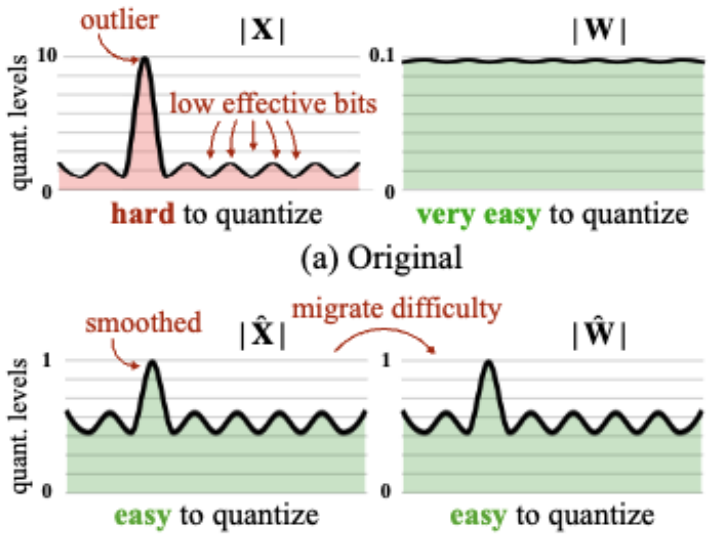
채널 별로, activation 변동성을 weigh로 얼마나 넘겨줄지 결정

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023, July). Smoothquant: Accurate and efficient post-training quantization for large language models. In International conference on machine learning (pp. 38087-38099). PMLR.

Activation Outliers

❖ SmoothQuant (PMLR'23)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 변동성이 weight 대비 크다 → Activation Quantization 난이도 ↑
- ➢ 핵심 아이디어: Activation 변동성 일부를 weight로 이전



$$Y = WX$$

$$Y = (W \cdot S^{-1}) \cdot (S \cdot X) = \hat{W}\hat{X}$$

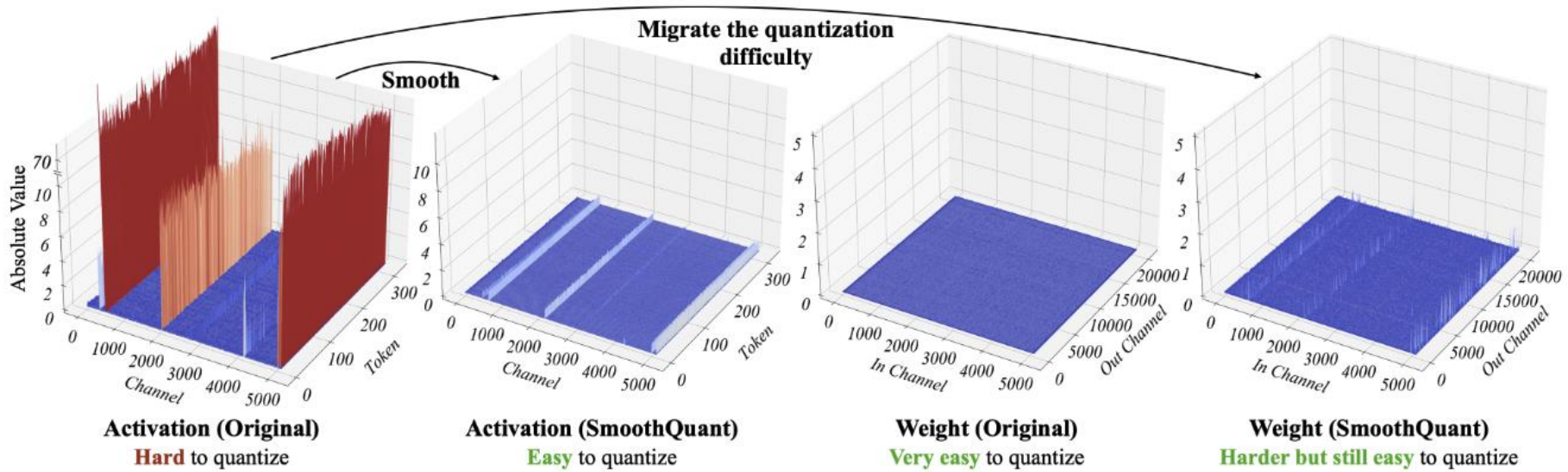
변동성 이전 전후 동일

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023, July). Smoothquant: Accurate and efficient post-training quantization for large language models. In International conference on machine learning (pp. 38087-38099). PMLR.

Activation Outliers

❖ SmoothQuant (PMLR'23)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 변동성이 weight 대비 크다 → Activation Quantization 난이도 ↑
- ➢ 핵심 아이디어: Activation 변동성 일부를 weight로 이전



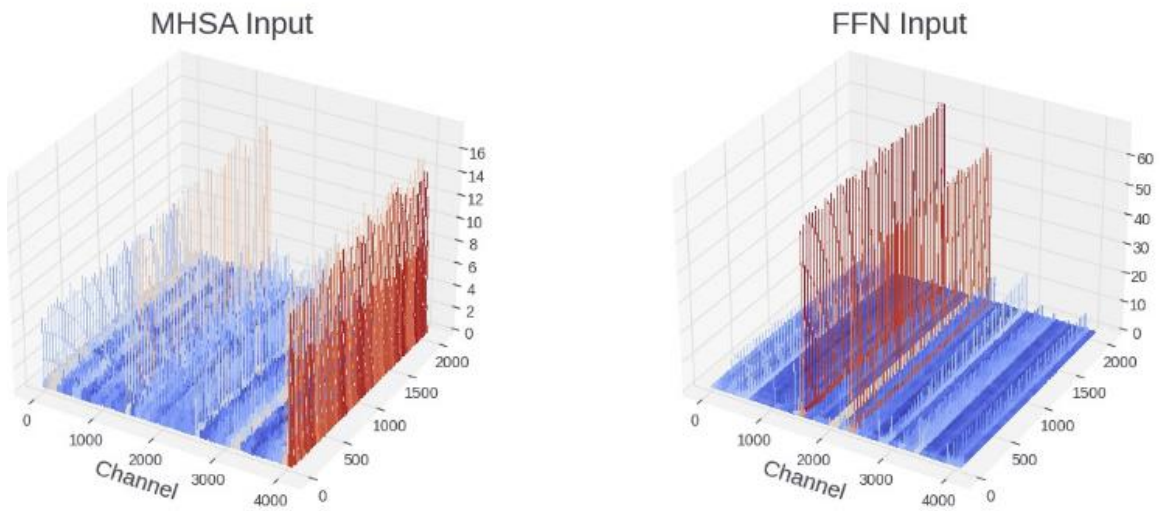
Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023, July). Smoothquant: Accurate and efficient post-training quantization for large language models. In International conference on machine learning (pp. 38087-38099). PMLR.

Activation Outliers

❖ SpinQuant (ICLR'25)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 outlier는 특정 채널에 분포
- ➢ 핵심 아이디어: 회전 행렬을 통해 quantization 친화적 분포로 변형

LLaMa-2 7B Activation Distribution

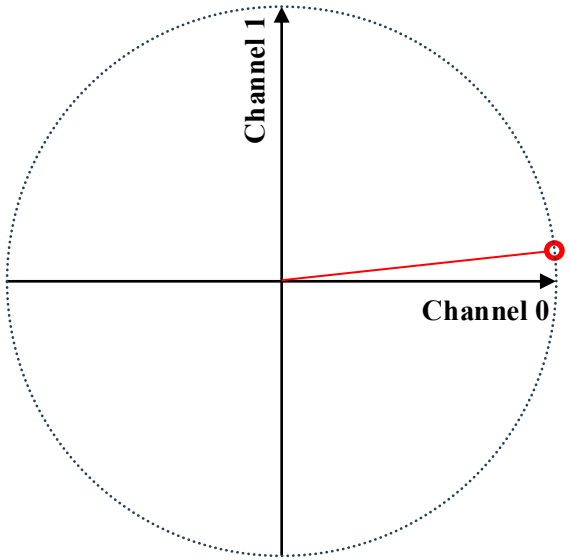


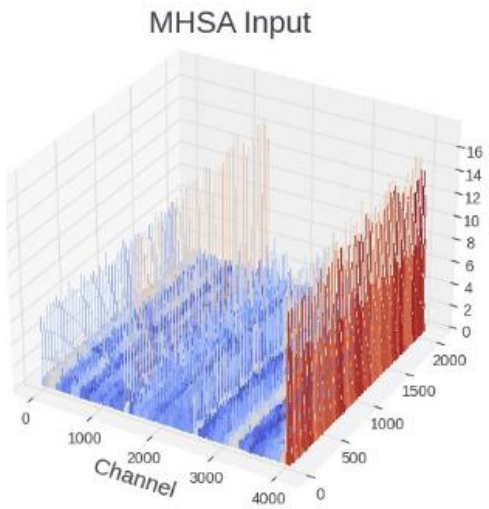
Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., ... & Blankevoort, T. SpinQuant: LLM Quantization with Learned Rotations. In The Thirteenth International Conference on Learning Representations.

Activation Outliers

❖ SpinQuant (ICLR'25)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 outlier는 특정 채널에 분포
- ➢ 핵심 아이디어: 회전 행렬을 통해 quantization 친화적 분포로 변형



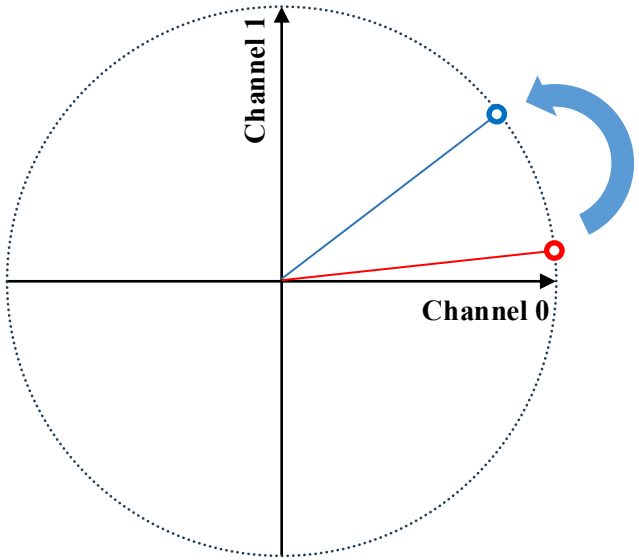


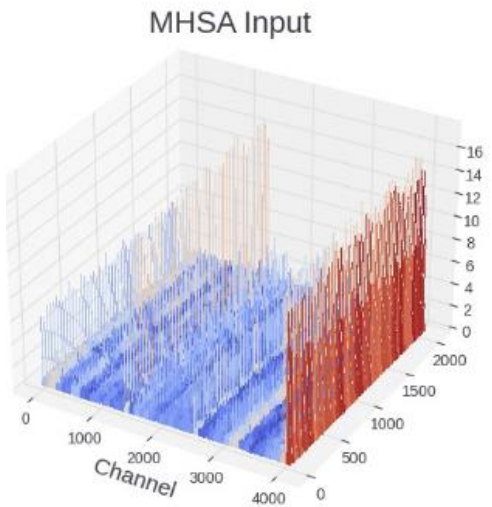
Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., ... & Blankevoort, T. SpinQuant: LLM Quantization with Learned Rotations. In The Thirteenth International Conference on Learning Representations.

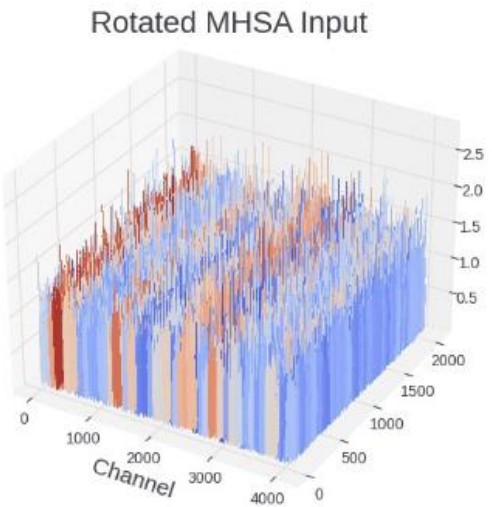
Activation Outliers

❖ SpinQuant (ICLR'25)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 outlier는 특정 채널에 분포
- ➢ 핵심 아이디어: 회전 행렬을 통해 quantization 친화적 분포로 변형





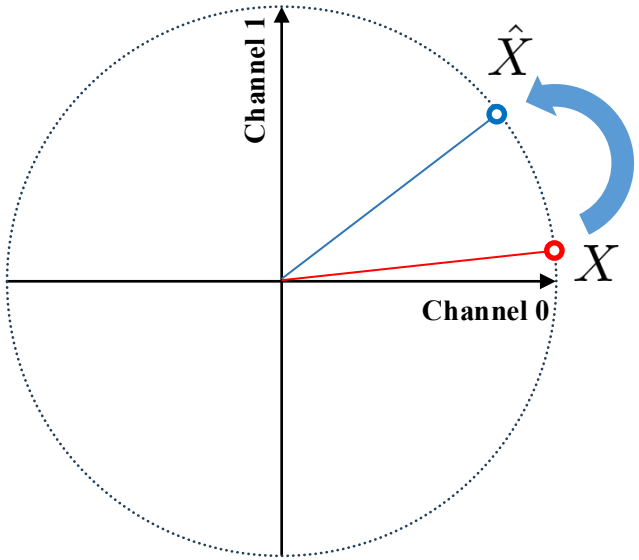


Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., ... & Blankevoort, T. SpinQuant: LLM Quantization with Learned Rotations. In The Thirteenth International Conference on Learning Representations.

Activation Outliers

❖ SpinQuant (ICLR'25)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 outlier는 특정 채널에 분포
- ➢ 핵심 아이디어: 회전 행렬을 통해 quantization 친화적 분포로 변형



$$Y = (W \cdot R^T) \cdot (R \cdot X) = \hat{W} \cdot \hat{X}$$

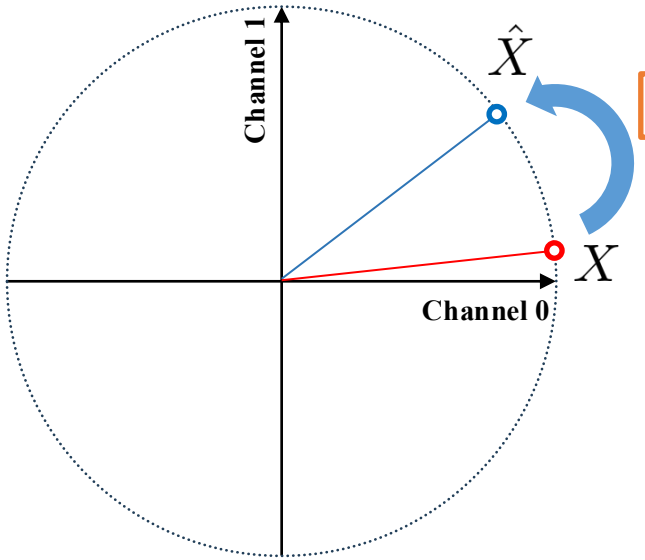
변동성 이전 전후 동일

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., ... & Blankevoort, T. SpinQuant: LLM Quantization with Learned Rotations. In The Thirteenth International Conference on Learning Representations.

Activation Outliers

❖ SpinQuant (ICLR'25)

- ➢ PTQ (Static) 계열
- ➢ 가정: Activation의 outlier는 특정 채널에 분포
- ➢ 핵심 아이디어: 회전 행렬을 통해 quantization 친화적 분포로 변형



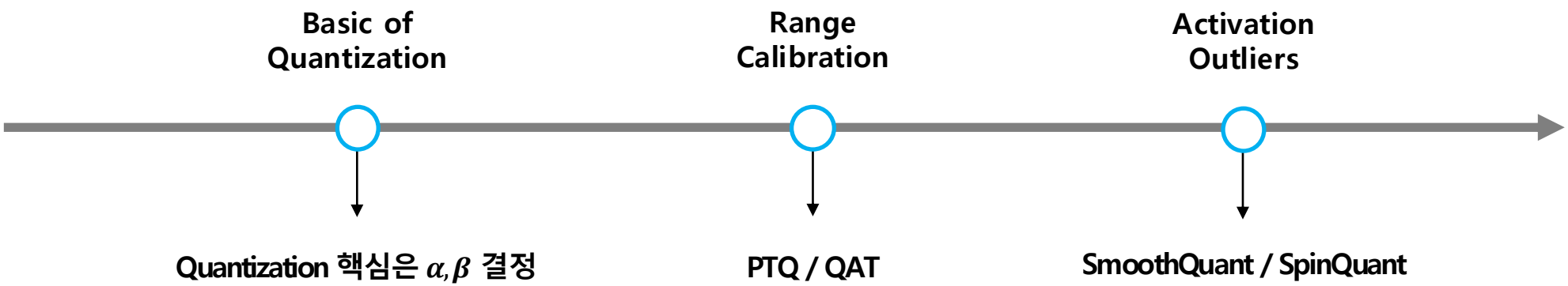
R PTQ calibration 시 학습

$$Y = (W \cdot R^T) \cdot (R \cdot X) = \hat{W} \cdot \hat{X}$$

변동성 이전 전후 동일

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., ... & Blankevoort, T. SpinQuant: LLM Quantization with Learned Rotations. In The Thirteenth International Conference on Learning Representations.

Summary


Basic of Quantization
Range Calibration
Activation Outliers
Quantization 핵심은 α, β 결정
PTQ / QAT
SmoothQuant / SpinQuant

고맙습니다

Data Mining
Quality Analytics

Reference

- Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficientneural network inference. In Low-power computer vision (pp. 291-326). Chapman and Hall/CRC.
- Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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. 27042713).
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
- Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). 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. 27042713).
- Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023, July). Smoothquant: Accurate and efficient post-training quantization for large language models. In International conference on machine learning (pp. 38087-38099). PMLR.
- Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., ... &Blankevoort, T. SpinQuant: LLM Quantization with Learned Rotations. In The Thirteenth International Conference on Learning Representations.
- https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization