AIMET

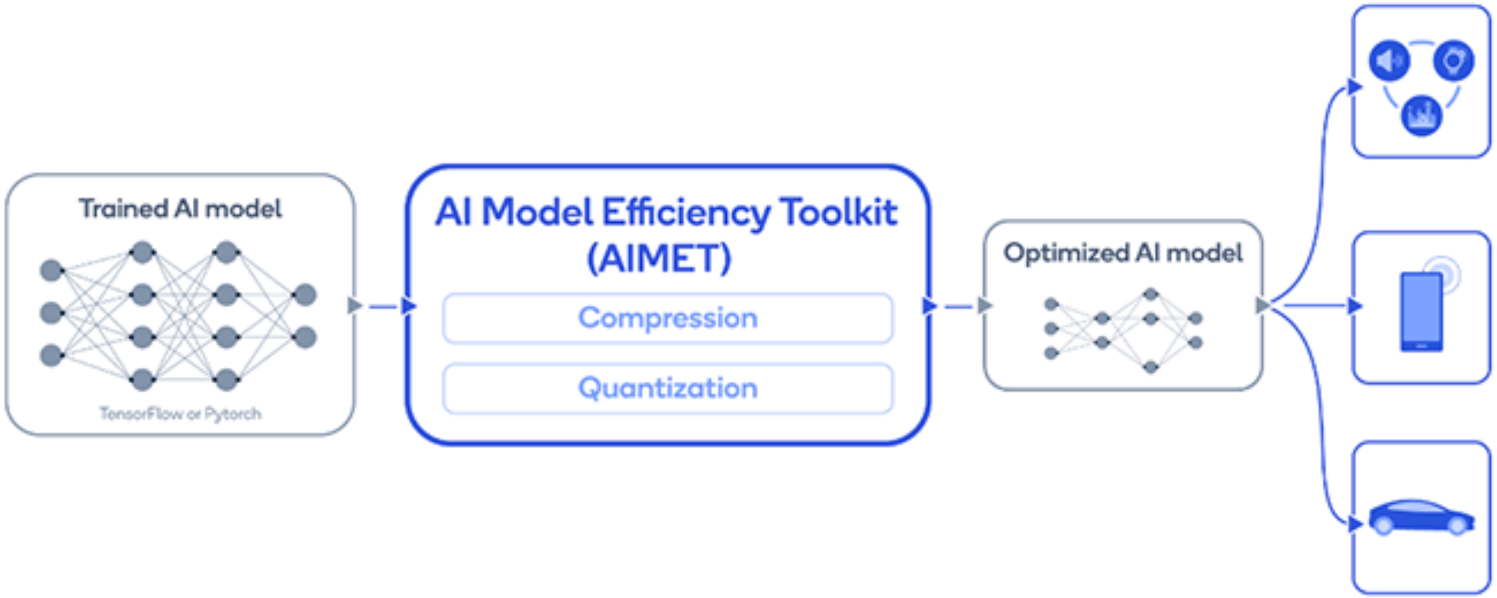
presenter: Jung Tae-young

tee.ty.jung@openedges.com

https://github.com/Tee0125

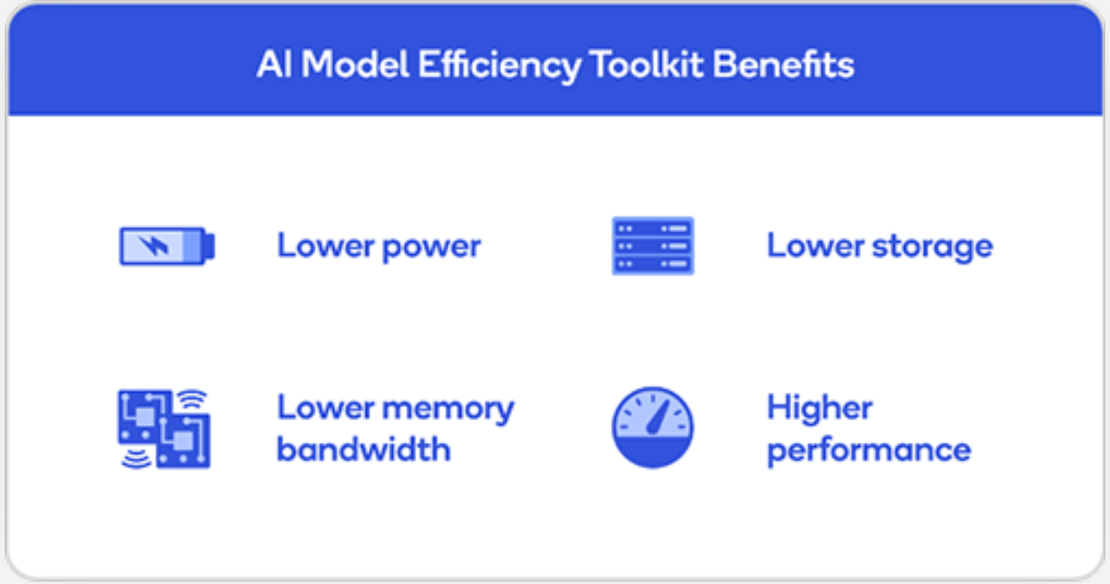
AIMET

- AI Model Efficiency Toolkit developed by Qualcomm



AIMET

- AI Model Efficiency Toolkit developed by Qualcomm



AIMET :: Supported Features

- Quantization
- Model Compression
- Visualization

AIMET :: Supported Features

- Quantization
  - Cross-Layer Equalization
  - Bias Correction
  - Adaptive Rounding
  - Quantization Simulation
  - Quantization-aware Training
- Model Compression
- Visualization

AIMET :: Supported Features

- Quantization
- Model Compression
  - Spatial SVD
  - Channel Pruning
  - Per-layer compresion-ratio selection
- Visualization

AIMET :: Supported Features

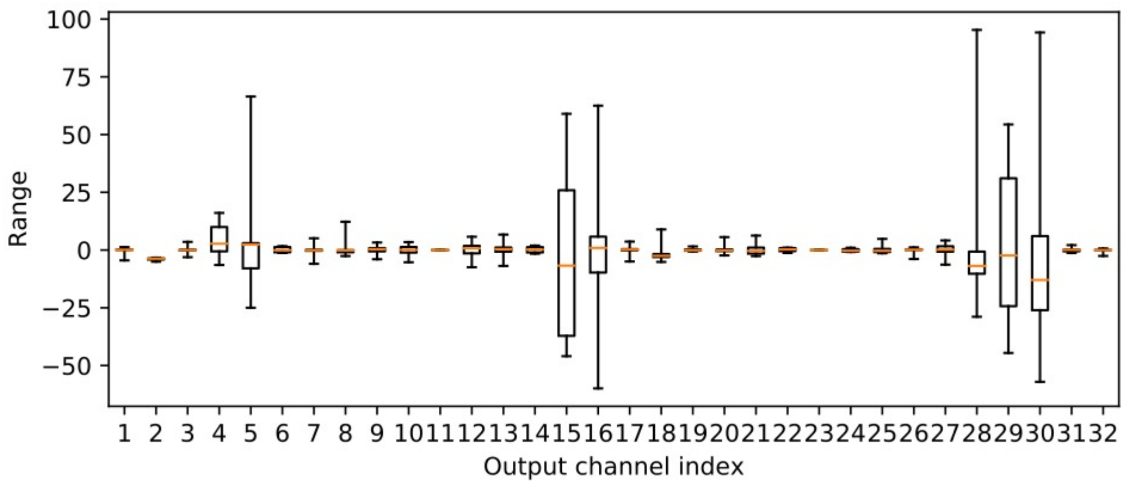
- Quantization
- Model Compression
- Visualization
  - Weight ranges
  - Per-layer compression sensitivity

AIMET :: Quantization Algorithms

- Cross-Layer Equalization
- Bias Correction
- AdaRounding

AIMET :: Cross-Layer Equaliazation

- Motivation



the weight distributions differ so strongly between output channels that the same set of quantization parameters cannot be used to quantize the full weight tensor effectively.

AIMET :: Cross-Layer Equaliazation

- Scaling Equivalence of Neural Network

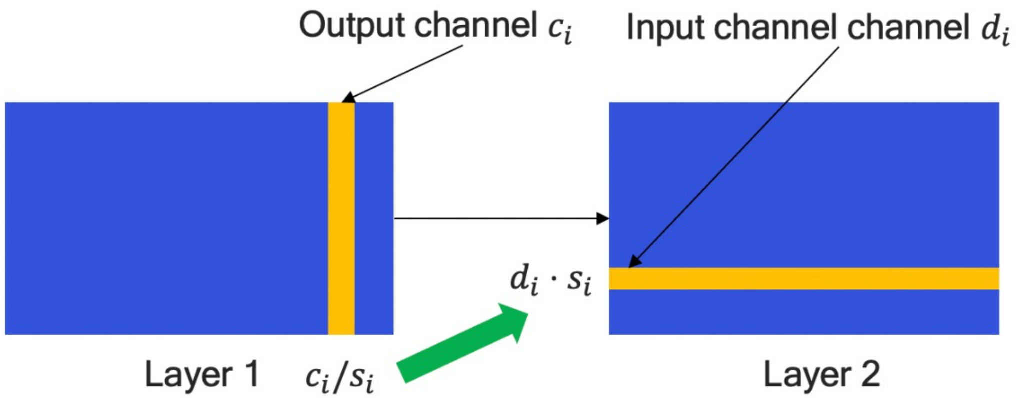


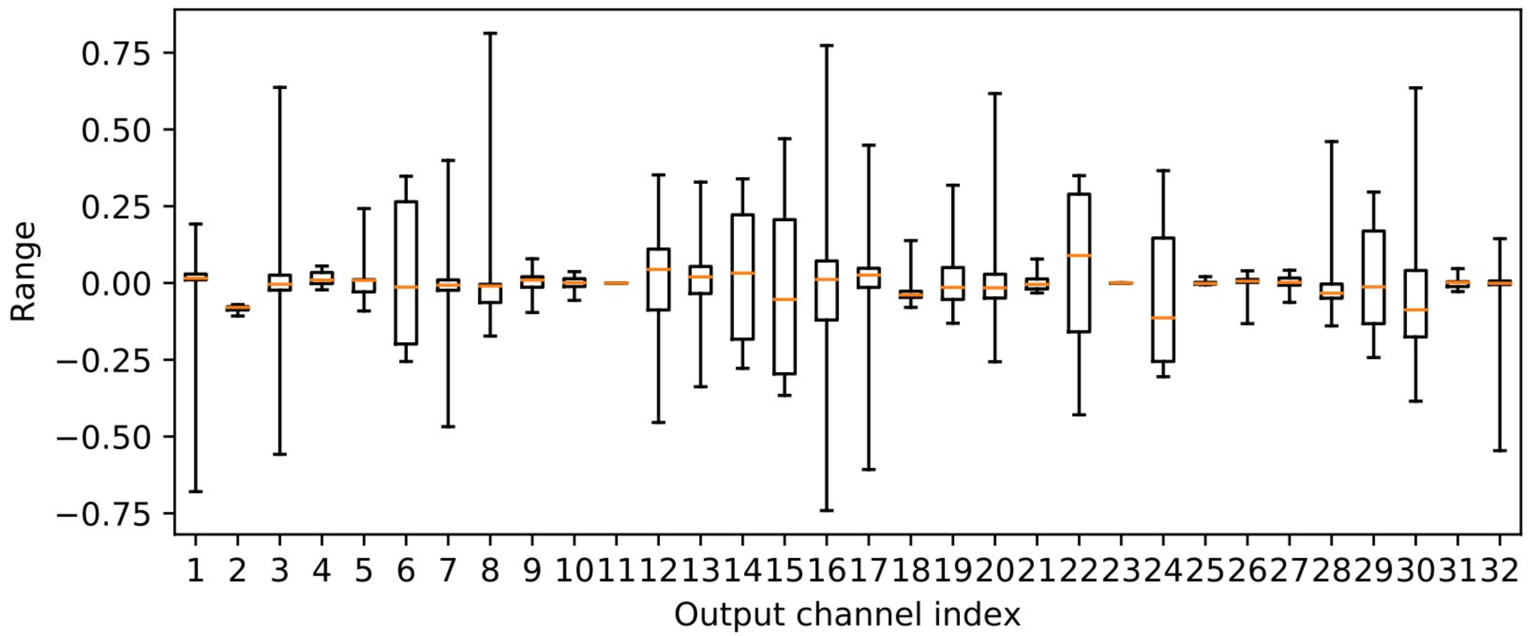
Figure 5. Illustration of the rescaling for a single channel. If scaling factor s_i scales c_i in layer 1; we can instead factor it out and multiply d_i in layer 2.

$$\mathbf{h} = f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \text{ and } \mathbf{y} = f(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)})$$

$$\begin{aligned}\mathbf{y} &= f(\mathbf{W}^{(2)} f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) \\ &= f(\mathbf{W}^{(2)}\mathbf{S}\hat{f}(\mathbf{S}^{-1}\mathbf{W}^{(1)}\mathbf{x} + \mathbf{S}^{-1}\mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) \\ &= f(\widehat{\mathbf{W}}^{(2)}\hat{f}(\widehat{\mathbf{W}}^{(1)}\mathbf{x} + \widehat{\mathbf{b}}^{(1)}) + \mathbf{b}^{(2)})\end{aligned}$$

AIMET :: Cross-Layer Equaliazation

- After Equalization…



AIMET :: Bias Correction

- Motivation
  - A commom assumption is that quantization error is unbiased and thus cancels out in a layer’s output
  - However, the quantization error on the weight might introduce biased error on the corresponding outputs.

AIMET :: Bias Correction

- Biased Quantization Error

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{W}}\mathbf{x}$$

$\widetilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\epsilon}\mathbf{x}$ where $\boldsymbol{\epsilon} = \widetilde{\mathbf{W}} - \mathbf{W}$

$$\begin{aligned}\mathbb{E}[\mathbf{y}] &= \mathbb{E}[\mathbf{y}] + \mathbb{E}[\boldsymbol{\epsilon}\mathbf{x}] - \mathbb{E}[\boldsymbol{\epsilon}\mathbf{x}] \\ &= \mathbb{E}[\widetilde{\mathbf{y}}] - \mathbb{E}[\boldsymbol{\epsilon}\mathbf{x}].\end{aligned}$$

$$\mathbb{E}\left[\boldsymbol{\epsilon}\mathbf{x}\right] = \boldsymbol{\epsilon}\mathbb{E}\left[\mathbf{x}\right]$$

AIMET :: Bias Correction

- After Bias Correction …

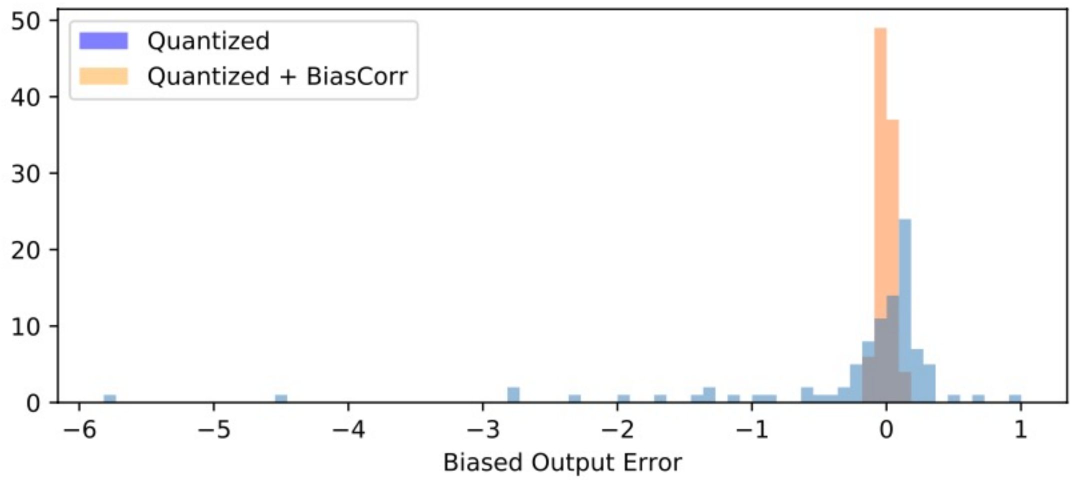


Figure 3. Per-channel biased output error introduced by weight quantization of the second depthwise-separable layer in MobileNetV2, before (blue) and after (orange) bias correction.

AIMET :: Quantization Algorithm

- Experimental Results

Model	FP32	INT8
Original model	72.94	41.40
DFQ (ours)	72.45	**72.33**
Per-channel quantization	72.94	71.44

Table 3. DeeplabV3+ (MobileNetV2 backend) on Pascal VOC segmentation challenge. Mean intersection over union (mIOU) evaluated at full precision and 8-bit integer quantized. Per-channel quantization is our own implementation of [16] applied post-training.

Model	FP32	INT8
Original model	68.47	10.63
DFQ (ours)	68.56	**67.91**
Per-channel quantization	68.47	67.52

Table 4. MobileNetV2 SSD-lite on Pascal VOC object detection challange. Mean average precision (mAP) evaluated at full precision and 8-bit integer quantized. Per-channel quantization is our own implementation of [16] applied post-training.

AIMET :: Compression Algorithms

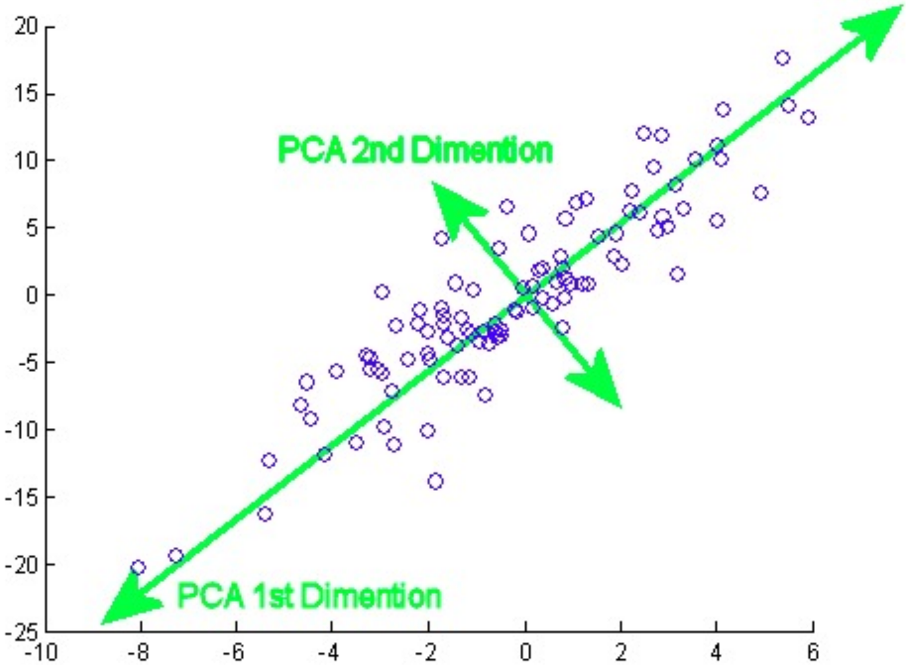
- Weight SVD
- Spatial SVD

AIMET :: Compression Algorithms

- Background
  - Principle Component Analysis
  - Singular Vector Decomposition

AIMET :: Compression Algorithms

- Principle Component Analysis



AIMET :: Compression Algorithms

- Principle Component Analysis (Eigen Face)

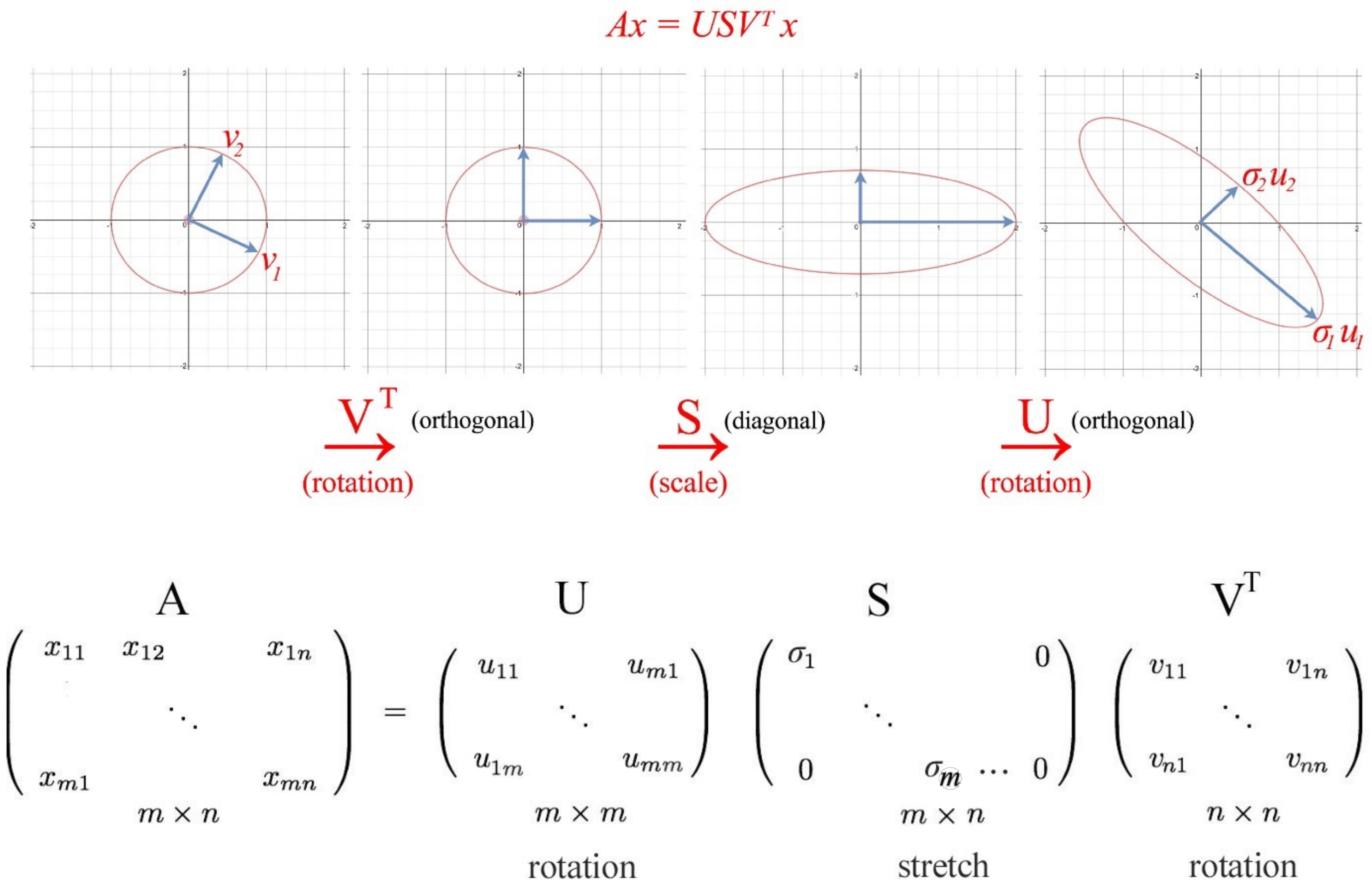
From: https://docs.opencv.org/3.4/da/d60/tutorial_face_main.html

AIMET :: Compression Algorithms

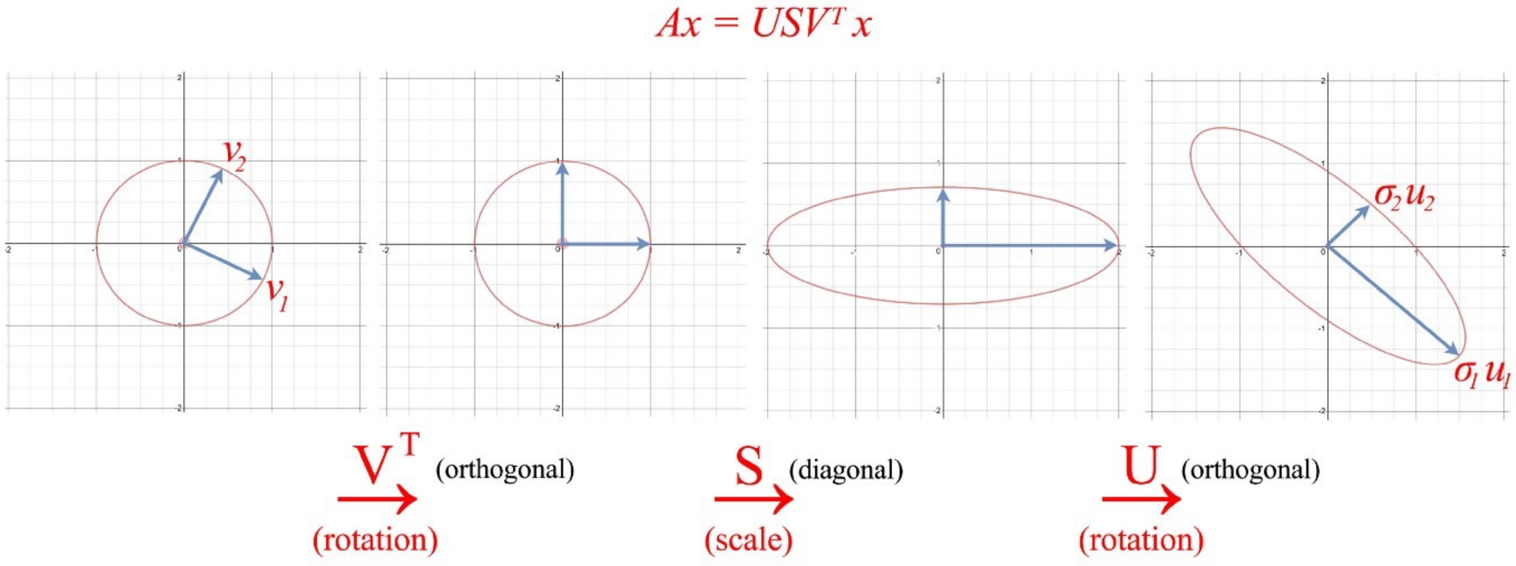
- Singular Vector Decomposition

$$A = U \cdot S \cdot V^T$$

AIMET :: Compression Algorithms



AIMET :: Compression Algorithms



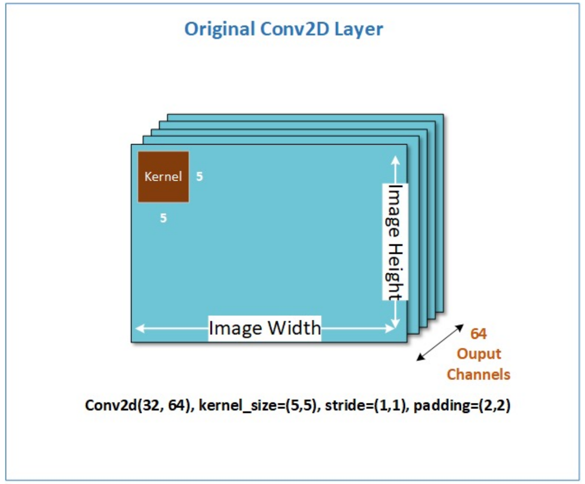
$$\underset{m \times n}{\overset{\mathrm{A}}{\begin{pmatrix} x_{11} & x_{12} & & x_{1n} \\ & & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} = \underset{\substack{m \times m \\ \text{rotation}}}{\overset{\mathrm{U}}{\begin{pmatrix} u_{11} & & u_{m1} \\ & \ddots & \\ u_{1m} & & u_{mm} \end{pmatrix}}} \underset{\substack{m \times n \\ \text{stretch}}}{\overset{\mathrm{S}}{\begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ 0 & & \sigma_m \cdots & 0 \end{pmatrix}}} \underset{\substack{n \times n \\ \text{rotation}}}{\overset{\mathrm{V^T}}{\begin{pmatrix} v_{11} & & v_{1n} \\ & \ddots & \\ v_{n1} & & v_{nn} \end{pmatrix}}}$$

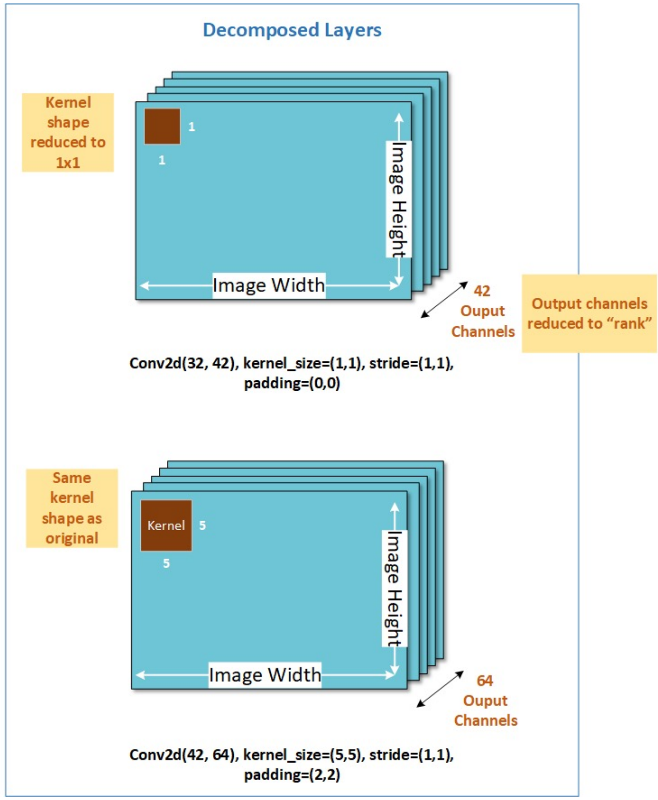
$$\overset{U}{\begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_m \\ | & & | \end{pmatrix}} \quad \overset{V}{\begin{pmatrix} | & & | \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \\ | & & | \end{pmatrix}}$$

eigenvectors for $\boldsymbol{AA^T}$ as u_i and $\boldsymbol{A^TA}$ as v_i

AIMET :: Compression Algorithms

- Weight SVD

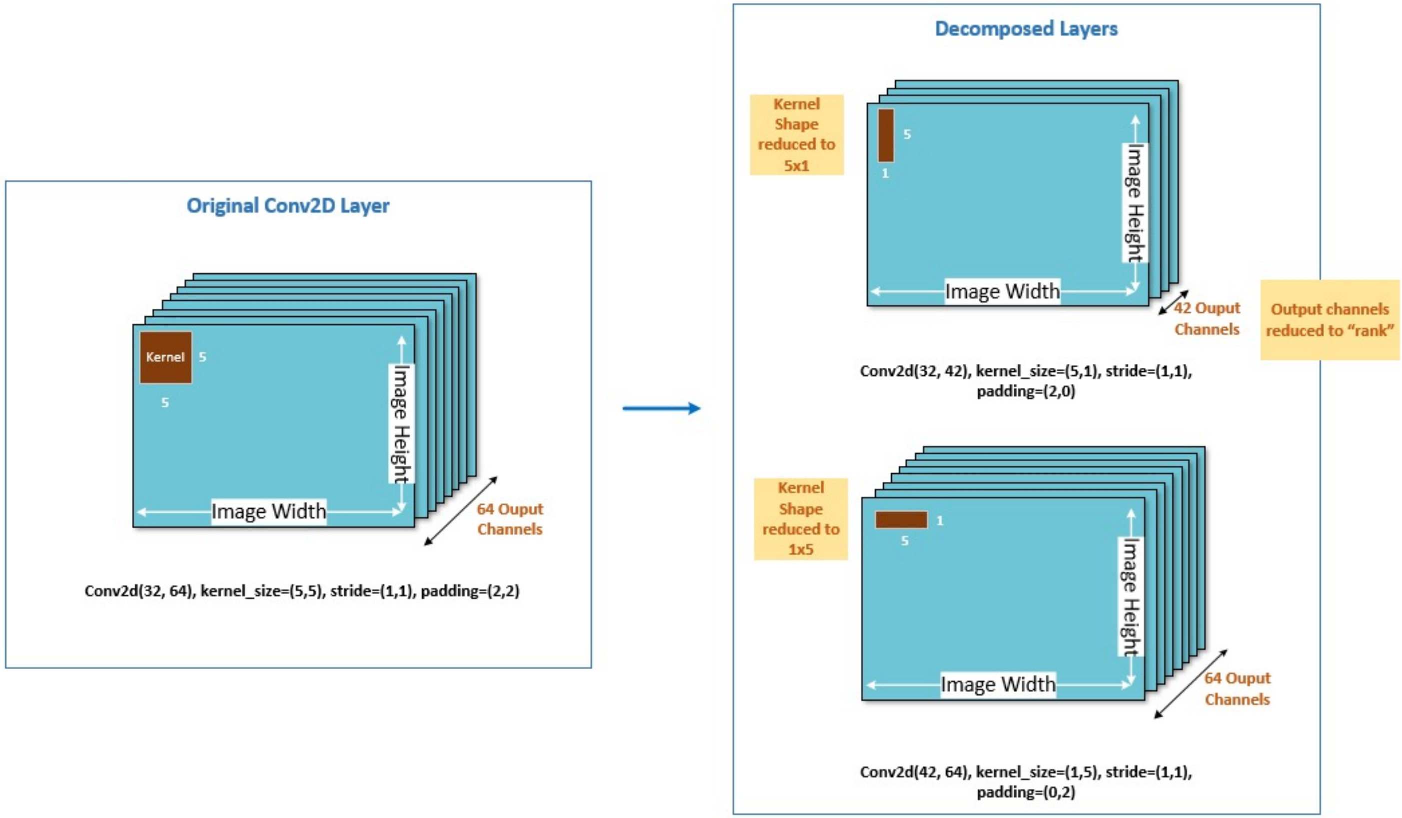




$$\underset{m \times m}{\overset{\mathrm{U}}{\begin{pmatrix} u_{11} & & u_{m1} \\ & \ddots & \\ u_{1m} & & u_{mm} \end{pmatrix}}} \quad \underset{m \times n}{\overset{\mathrm{S}}{\begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ & & & \ddots \\ 0 & & & 0 \end{pmatrix}}} \quad \underset{n \times n}{\overset{\mathrm{V}^{\mathrm{T}}}{\begin{pmatrix} v_{11} & & v_{1n} \\ & \ddots & \\ v_{n1} & & v_{nn} \end{pmatrix}}}$$

AIMET :: Compression Algorithms

- Spatial SVD



AIMET :: Compression Algorithms

Model Compression

AIMET can also significantly compress models. For popular models, such as Resnet-50 and Resnet-18, compression with spatial SVD plus channel pruning achieves 50% MAC (multiply-accumulate) reduction while retaining accuracy within approx. 1% of the original uncompressed model.

Models	Uncompressed model	50% Compressed model
ResNet18 (top1)	69.76%	68.56%
ResNet 50 (top1)	76.05%	75.75%

References

- Data-Free Quantization Through Weight Equalization and Bias Correction - https://arxiv.org/abs/1906.04721
- https://quic.github.io/aimet-pages/releases/1.16.2/user_guide/model_compression.html
- https://jonathan-hui.medium.com/machine-learning-singular-value-decomposition-svd-principal-component-analysis-pca-1d45e885e491