Mechanistic Interpretability 101

Episode 4

Einsum and Einops

Mechanistic Interpretability Tool Kit

Framework

PyTorch

TensorFlow

Visualization & Analysis Tools

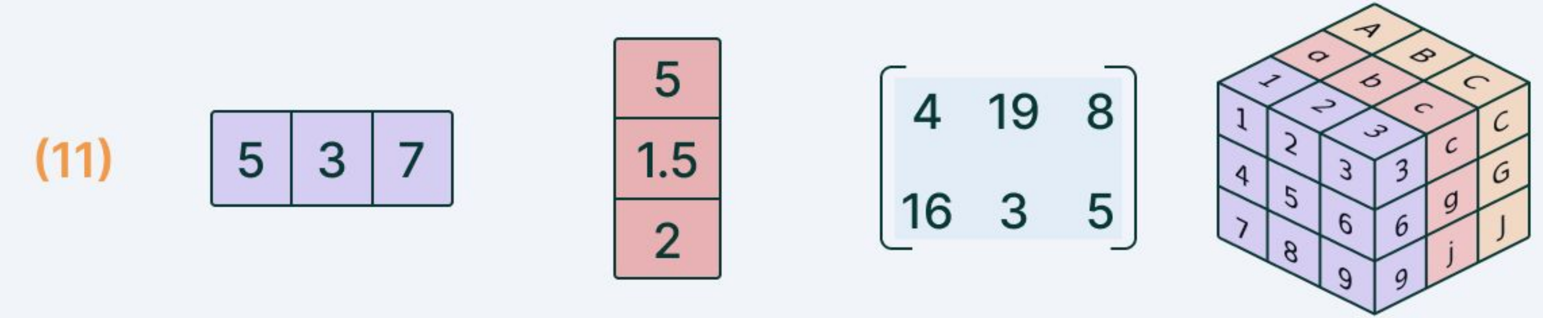
Weights & Biases

lucid

TransformerLens

Captum

Data



Scalar

Row Vector
(shape 1×3)

Column Vector
(shape 3×1)

Matrix

Tensor

Mechanistic Interpretability Tool Kit

Libraries