Memory-Efficient Neural Network Training for Wearable Devices

Maxim Nitsenko

DATA SCIENCE
UNIVERSITY OF PADOVA

DIPARTIMENTO
MATEMATICA

Introduction

Problem - High memory demands of neural network training, especially on constrained devices like smartwatches and health monitors.

Introduction

Problem - High memory demands of neural network training, especially on constrained devices like smartwatches and health monitors.

Motivation - On device training allows personalization, data transfer savings, and preserves privacy.

Introduction

Problem - High memory demands of neural network training, especially on constrained devices like smartwatches and health monitors.

Motivation - On device training allows personalization, data transfer savings, and preserves privacy.

Project Aim - Develop techniques to reduce peak training memory without large accuracy loss.

Introduction

Problem - High memory demands of neural network training, especially on constrained devices like smartwatches and health monitors.

Motivation - On device training allows personalization, data transfer savings, and preserves privacy.

Project Aim - Develop techniques to reduce peak training memory without large accuracy loss.

Result - 82× peak memory reduction (9.3 GB → 113 MB) with 3.3% accuracy drop (98 % → 95%).

Types of Memory

Model Memory - Stores model parameters

Types of Memory

Model Memory - Stores model parameters

Optimizer Memory - Stores gradients and momentum buffers

Types of Memory

Model Memory - Stores model parameters

Optimizer Memory - Stores gradients and momentum buffers

Activation Memory - Stores intermediate outputs

Types of Memory

Model Memory - Stores model parameters

Optimizer Memory - Stores gradients and momentum buffers

Activation Memory - Stores intermediate outputs

Other - Framework overhead

Optimization Stack

Configuration 1 (C1)

Resnet-50

Optimization Stack

Configuration 2 (C2)

Resnet-50
Dynamic Sparse Traing

Optimization Stack

Configuration 3 (C3)

Resnet-50
Dynamic Sparse Traing
Gradient Checkpointg

Optimization Stack

Configuration 4 (C4)

Resnet-50
Dynamic Sparse Traing
Gradient Checkpointg
Mictobatching

Optimization Stack

Configuration 5 (C5)

Dynamic Sparse Traing
Gradient Checkpointg
Mictobatching

Optimization Stack

Configuration 5 (C5)

Dynamic Sparse Traing
Gradient Checkpointg
Mictobatching
MobileNet

Optimization Stack

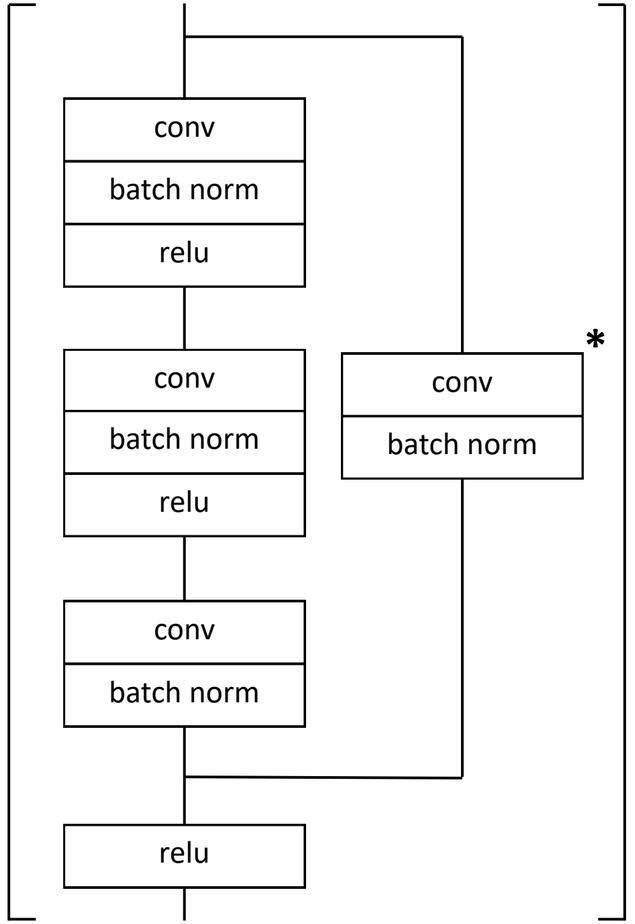
Configuration 6 (C6)

Dynamic Sparse Traing
Gradient Checkpointg
Mictobatching
MobileNet
Patching

Optimization Stack

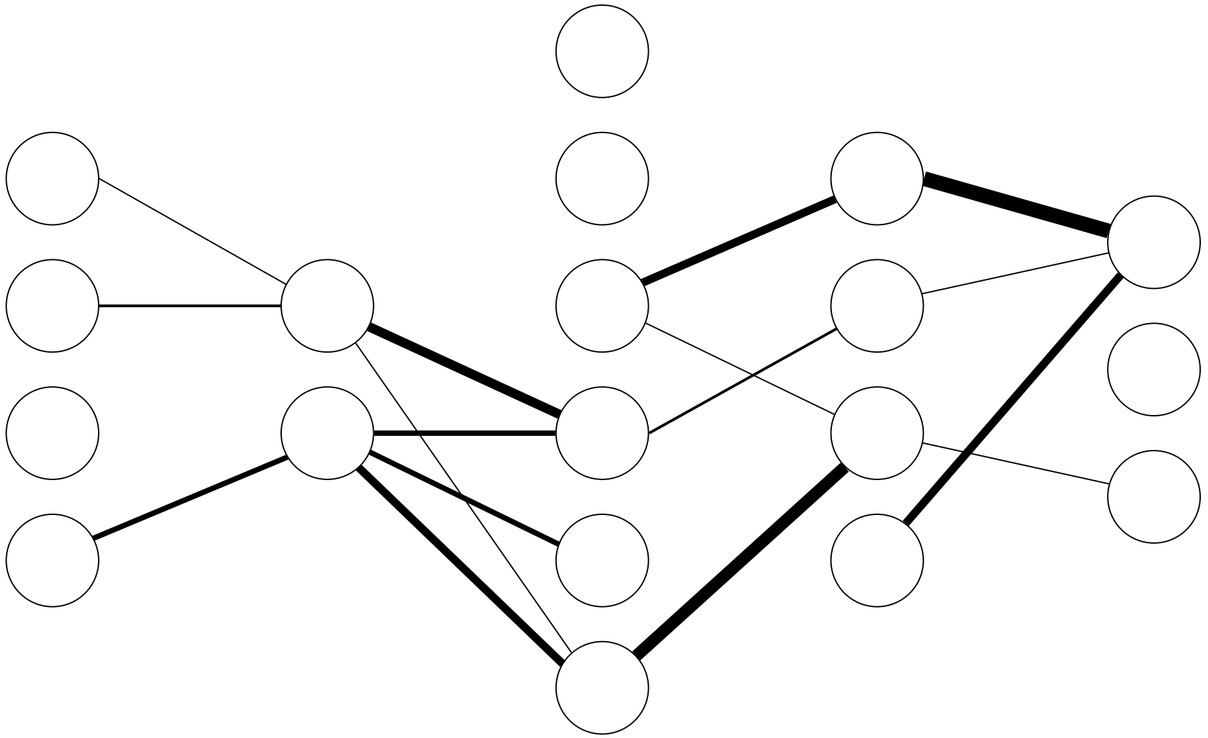
Resnet-50

Layer	Input Size	Output Size	Filter Shape; Stride
Convolution	$224 \times 224 \times 3$	$112 \times 112 \times 64$	$7, 7, 3, 64; 2$
Max Pool	$112 \times 112 \times 64$	$56 \times 56 \times 64$	$3, 3; 2$
Residual Stack 1	$56 \times 56 \times 64$	$56 \times 56 \times 256$	$\begin{bmatrix} 1, 1, 64 \vert 256, 64; 1 \\ 3, 3, 64, 64; 1 \\ 1, 1, 64, 256; 1 \end{bmatrix} \times 3$
Residual Stack 2	$56 \times 56 \times 256$	$28 \times 28 \times 512$	$\begin{bmatrix} 1, 1, 256 \vert 512, 128; 2 \vert 1 \\ 3, 3, 128, 128; 1 \\ 1, 1, 128, 512; 1 \end{bmatrix} \times 4$
Residual Stack 3	$28 \times 28 \times 512$	$14 \times 14 \times 1024$	$\begin{bmatrix} 1, 1, 512 \vert 1024, 256; 2 \vert 1 \\ 3, 3, 256, 256; 1 \\ 1, 1, 256, 1024; 1 \end{bmatrix} \times 6$
Residual Stack 4	$14 \times 14 \times 1024$	$7 \times 7 \times 2048$	$\begin{bmatrix} 1, 1, 1024 \vert 2048, 512; 2 \vert 1 \\ 3, 3, 512, 512; 1 \\ 1, 1, 512, 2048; 1 \end{bmatrix} \times 3$
Average Pool	$7 \times 7 \times 2048$	2048	$7, 7$
Fully Connected	2048	8	$2048, 8$



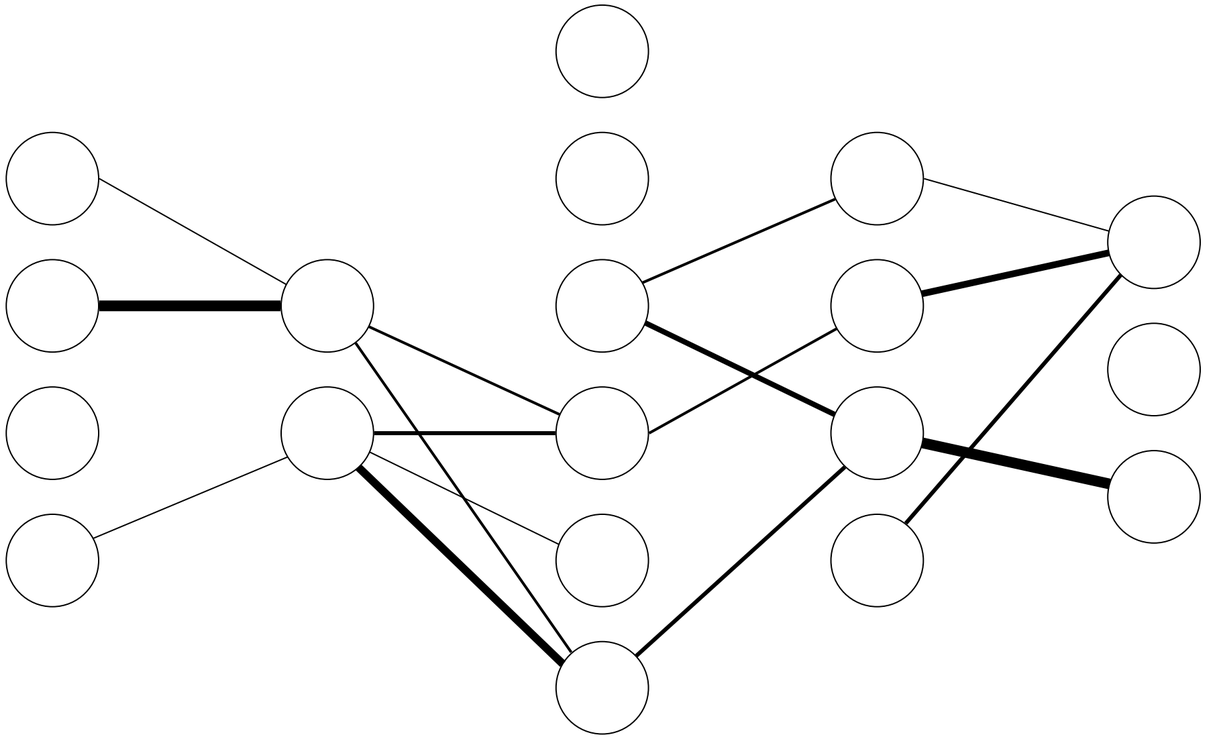
Optimization Stack

Dynamic Sparse Training



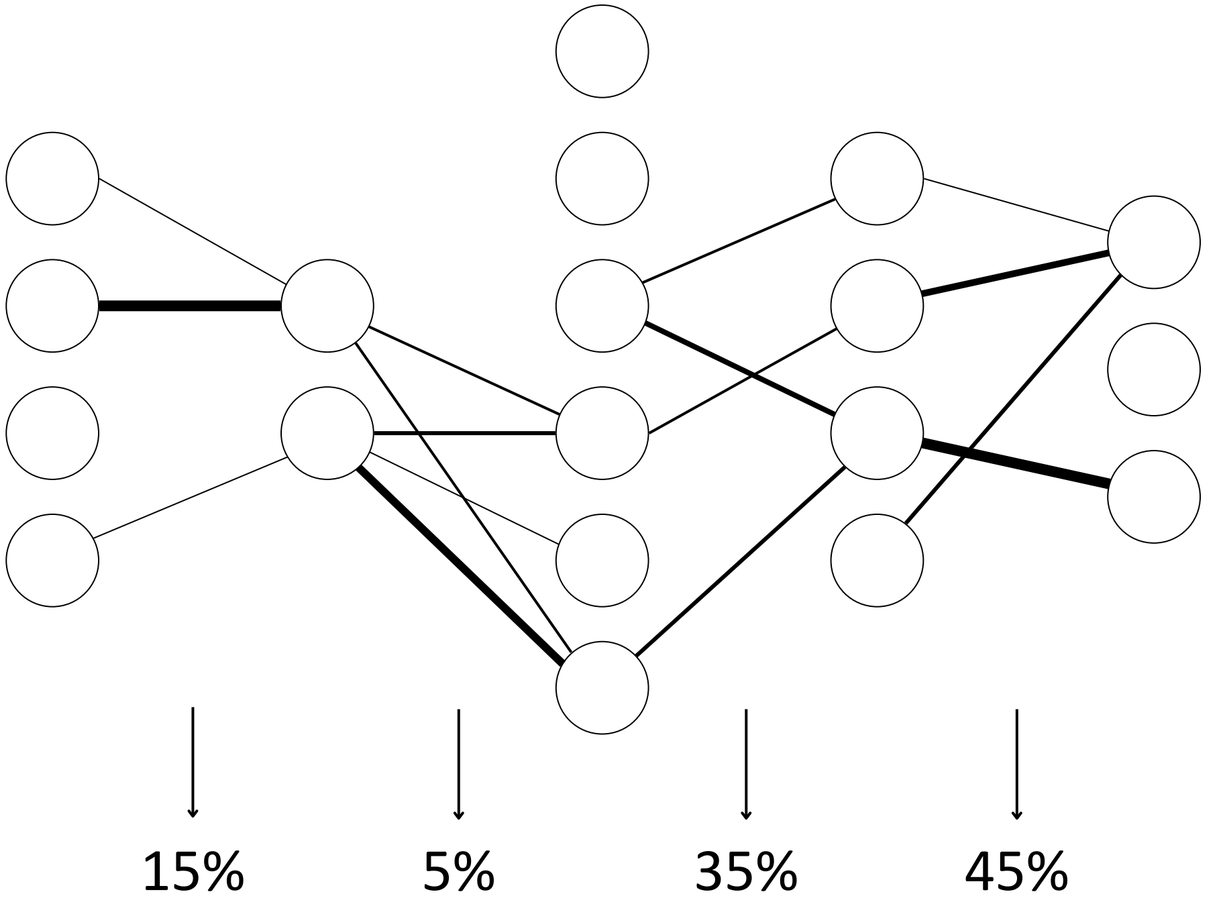
Optimization Stack

Dynamic Sparse Training



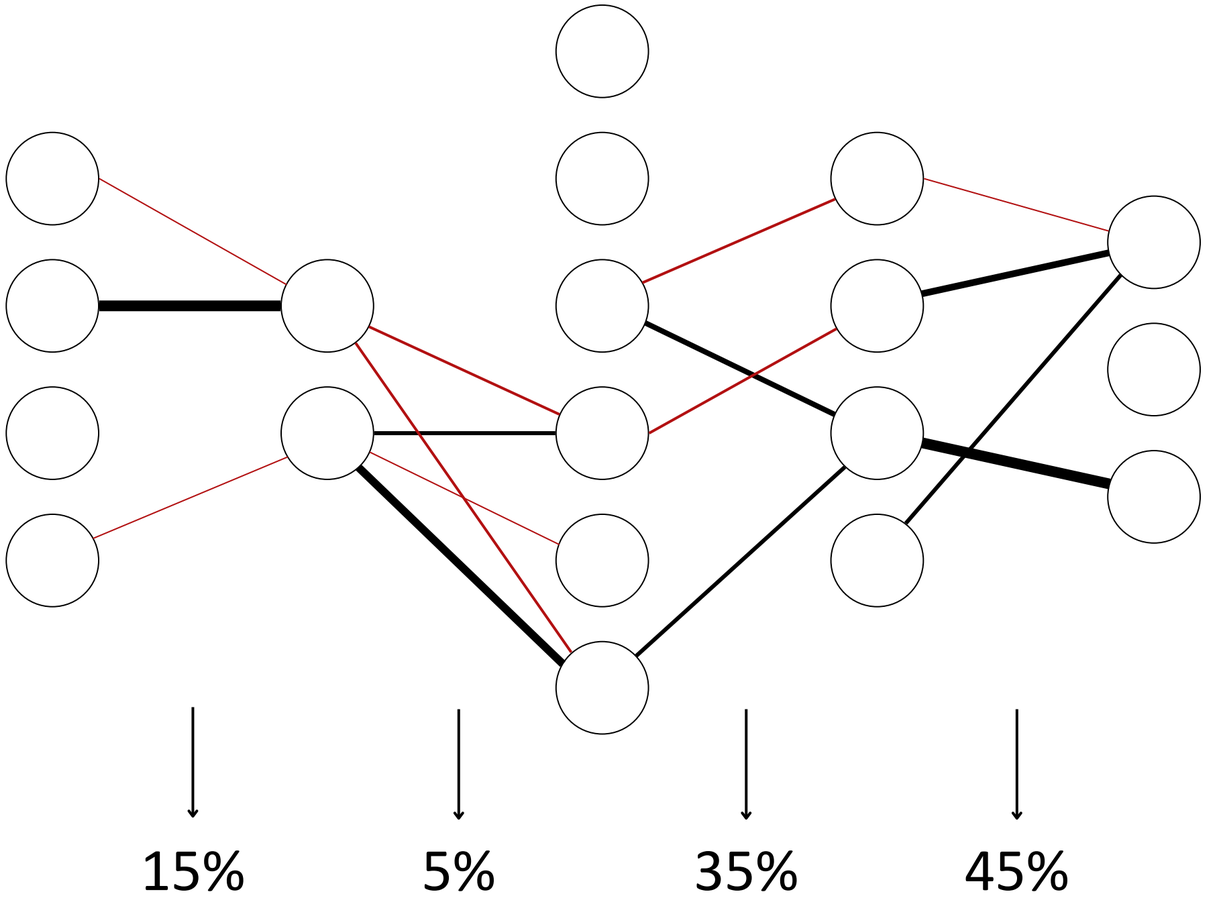
Optimization Stack

Dynamic Sparse Training



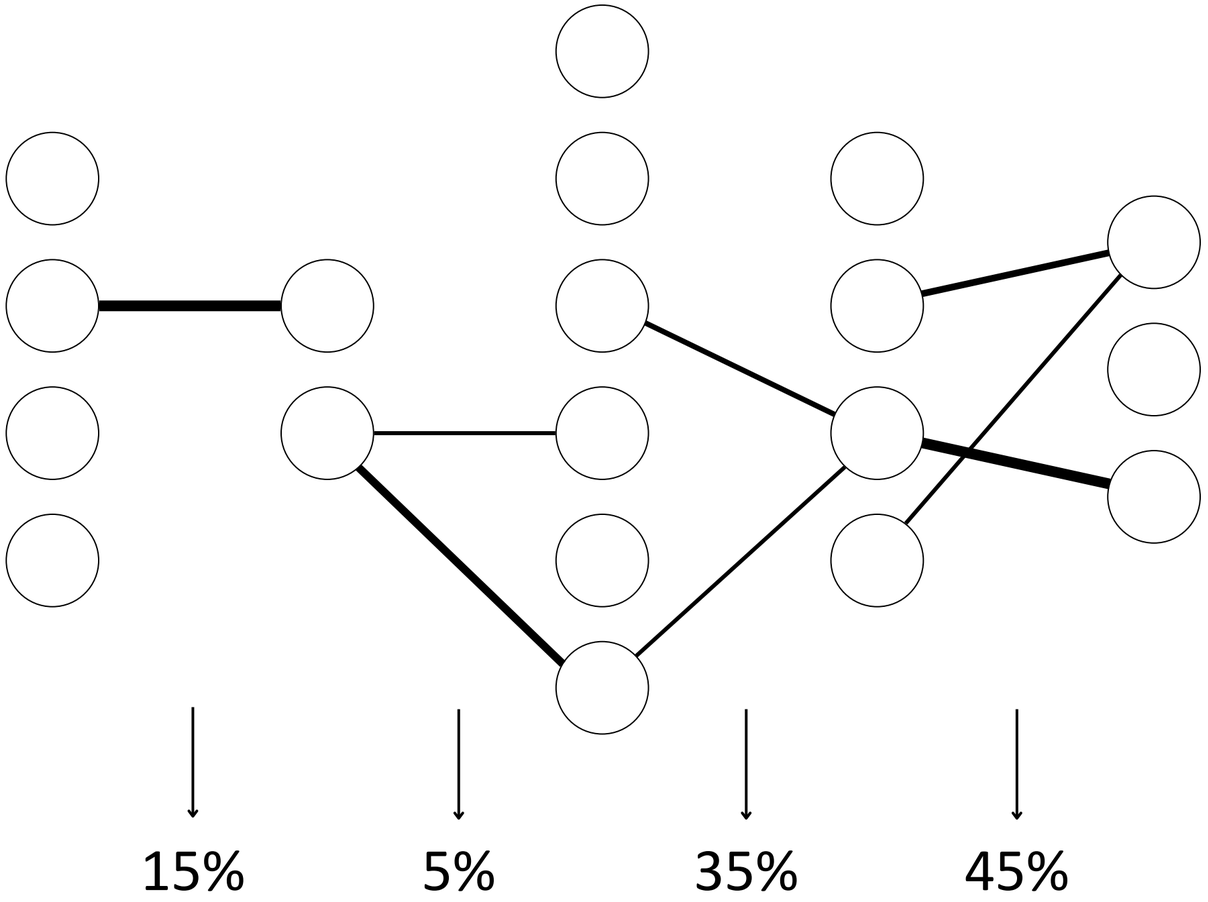
Optimization Stack

Dynamic Sparse Training



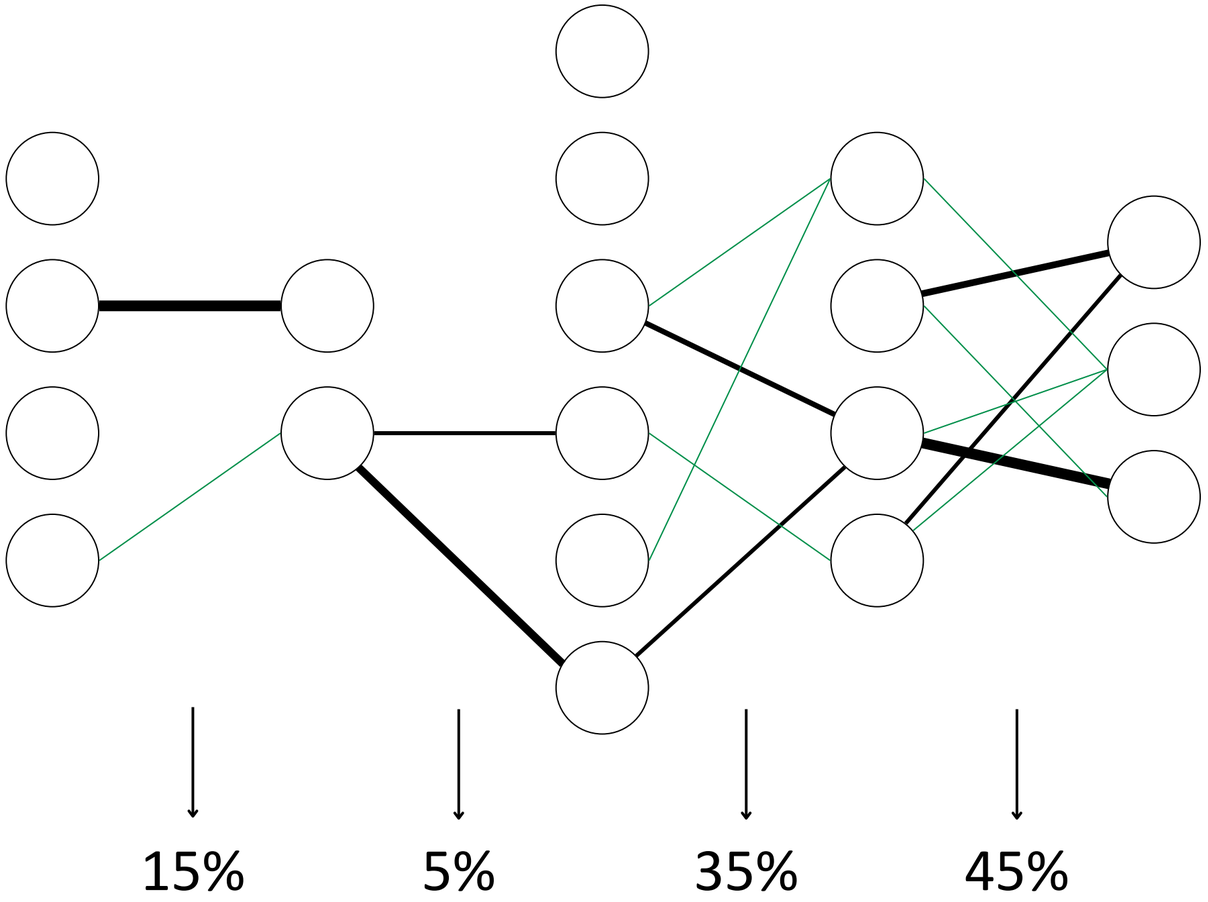
Optimization Stack

Dynamic Sparse Training



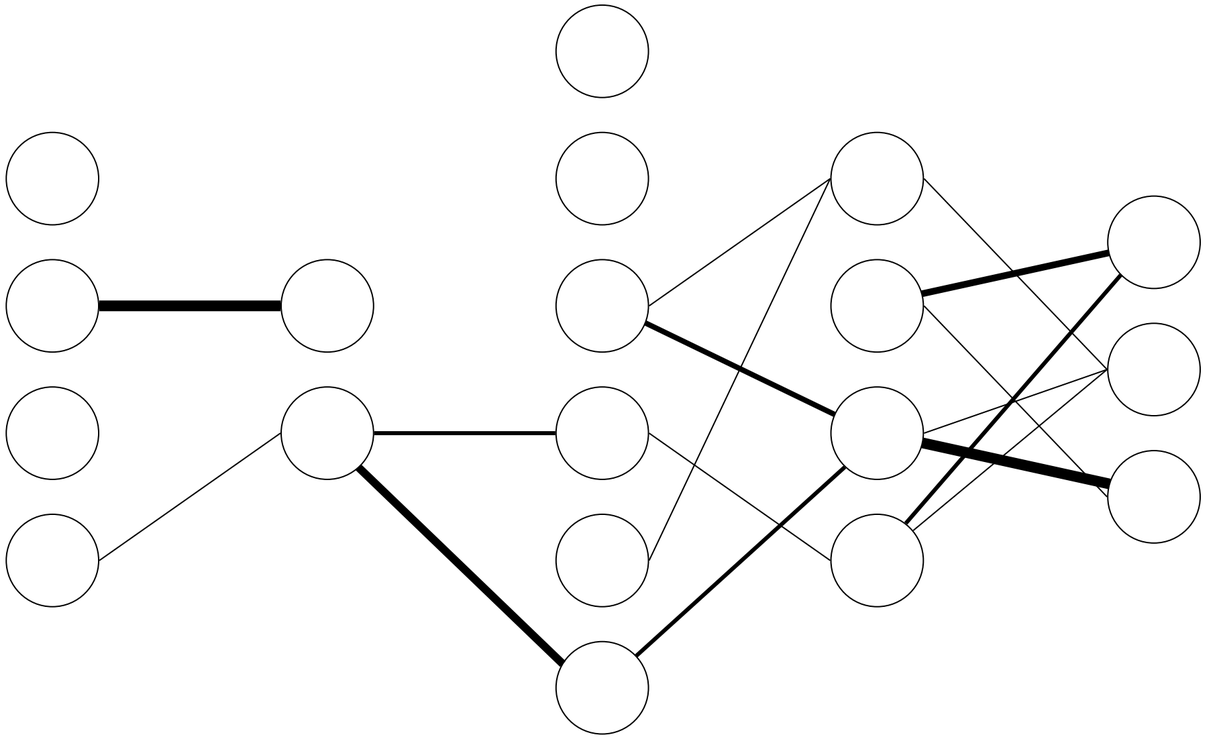
Optimization Stack

Dynamic Sparse Training



Optimization Stack

Dynamic Sparse Training

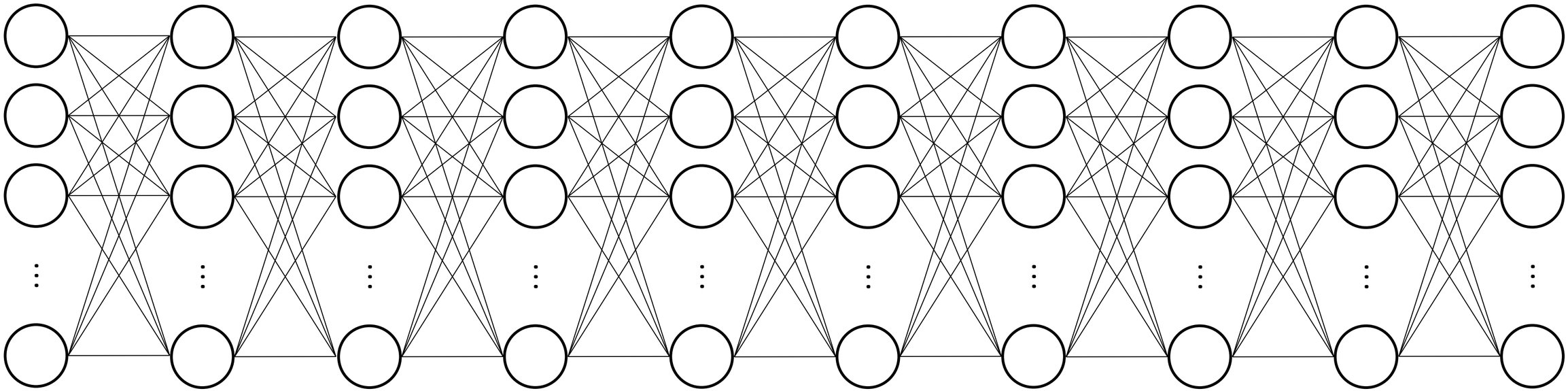


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} \, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)} \, w_{kj}^{(l+1)} \, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2 \left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

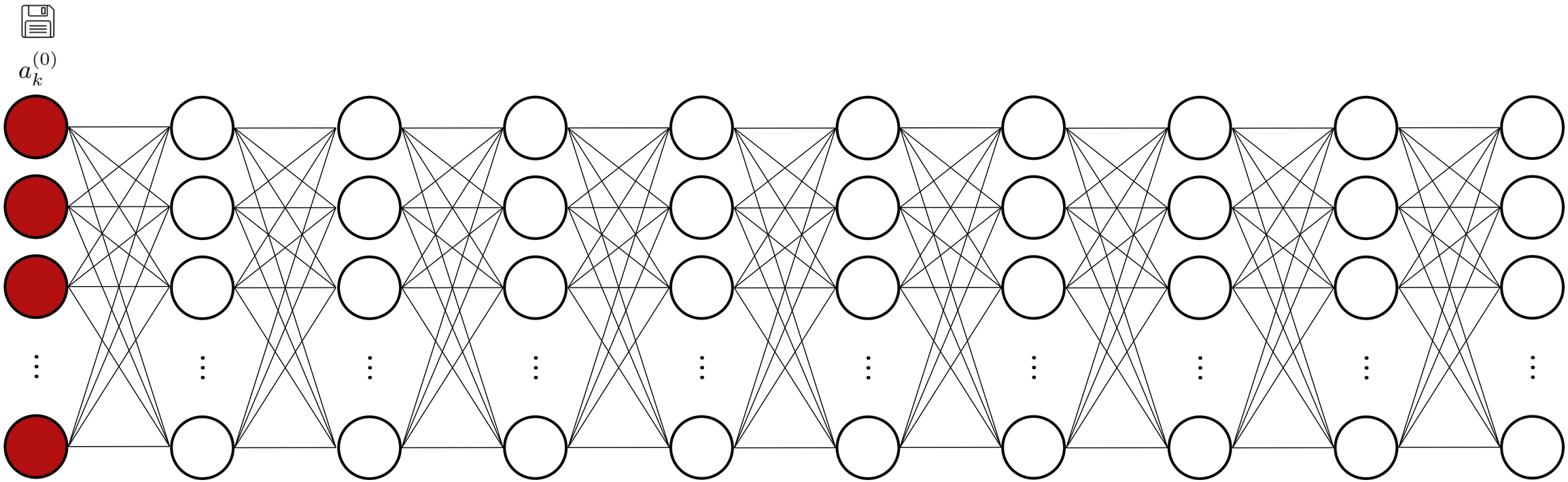


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

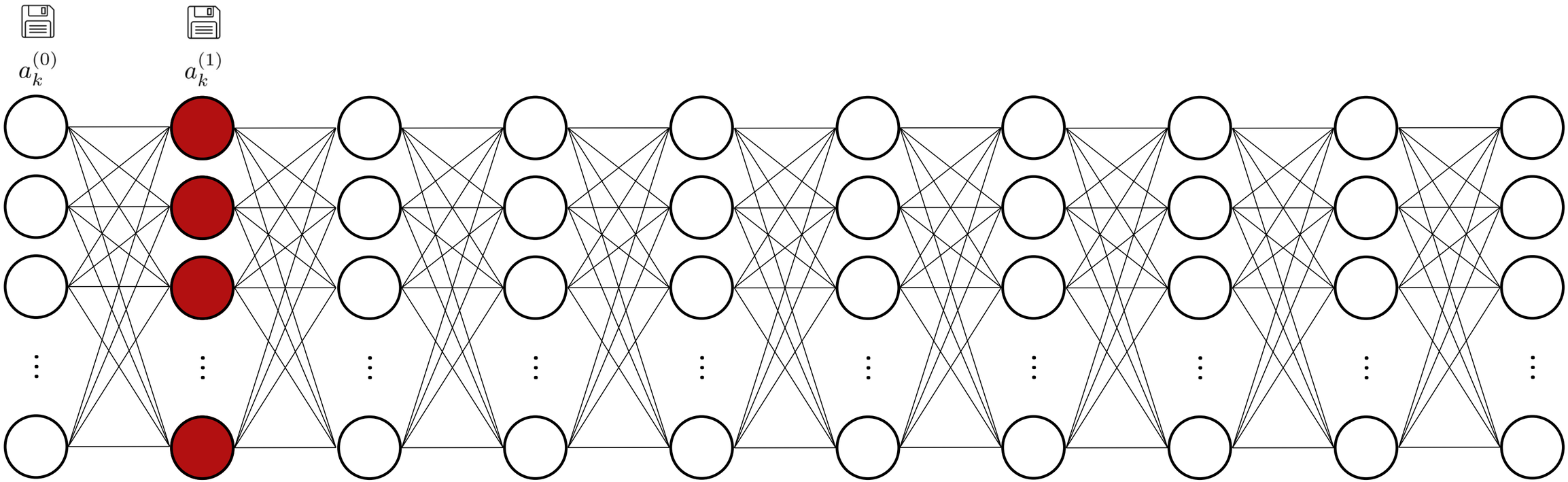


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

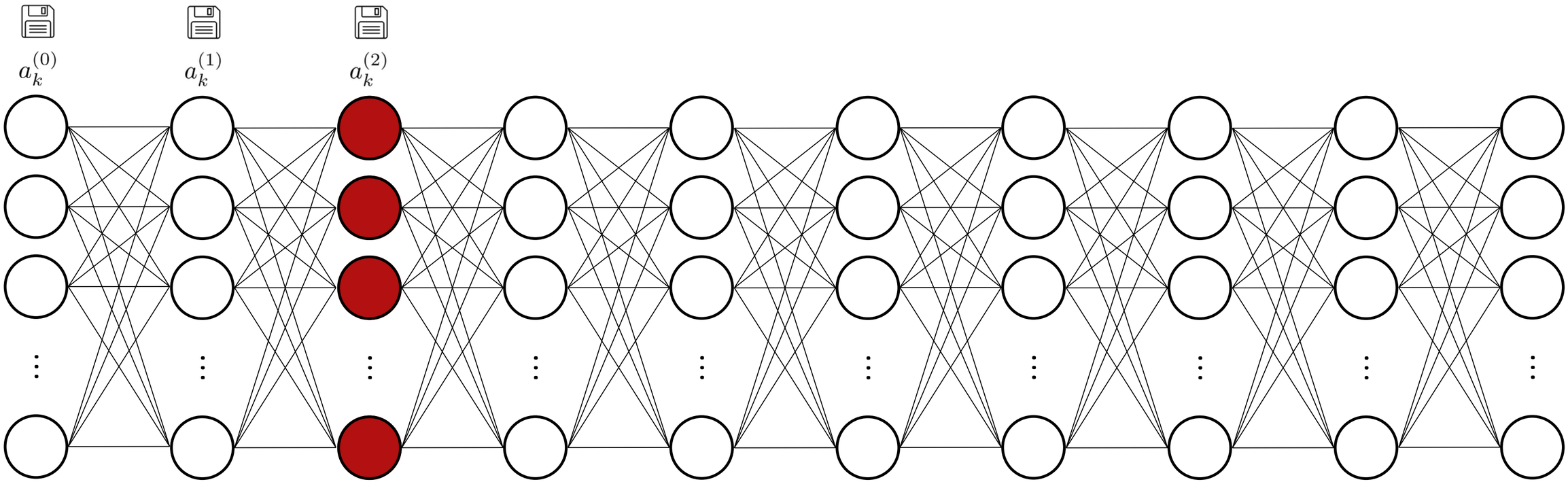


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

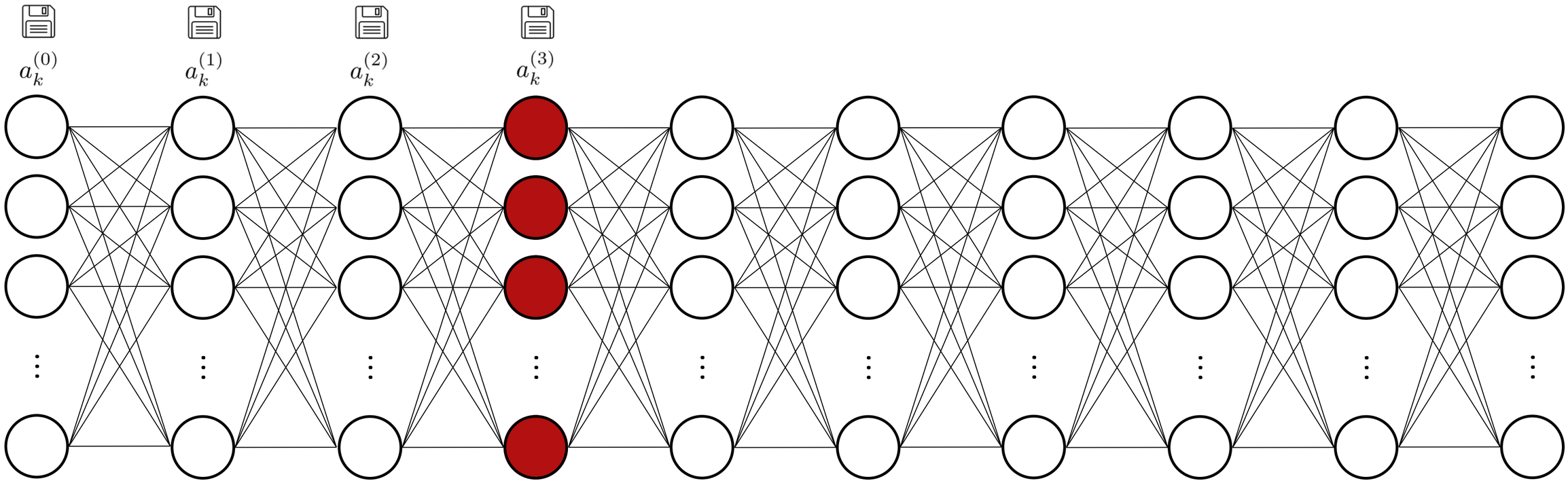


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

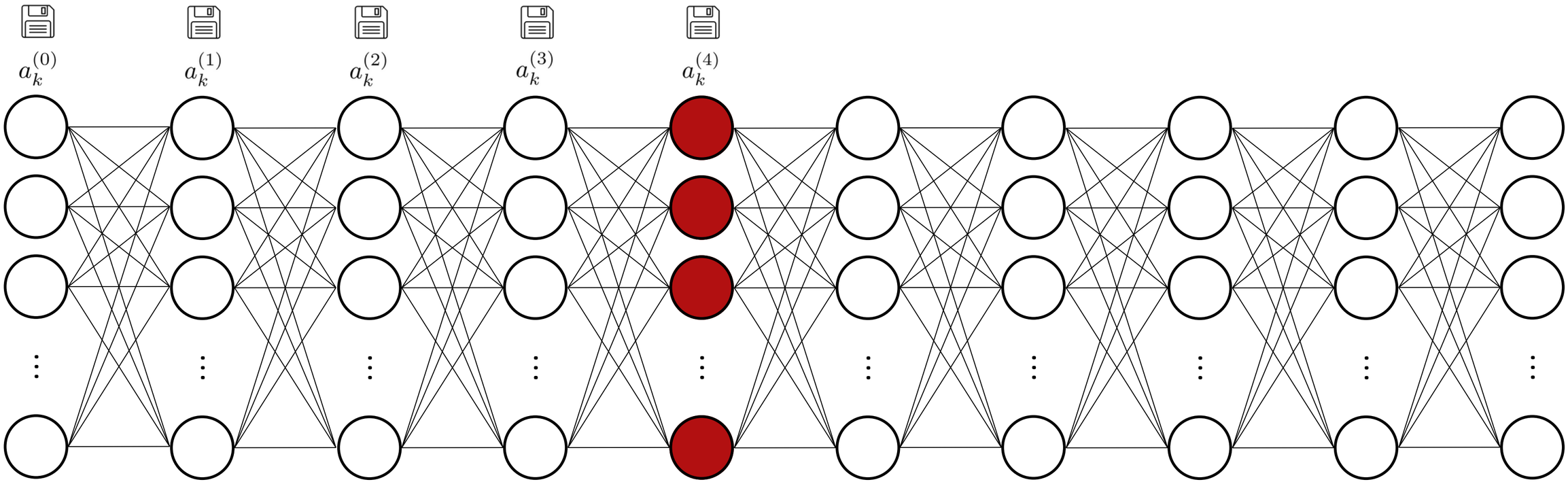


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

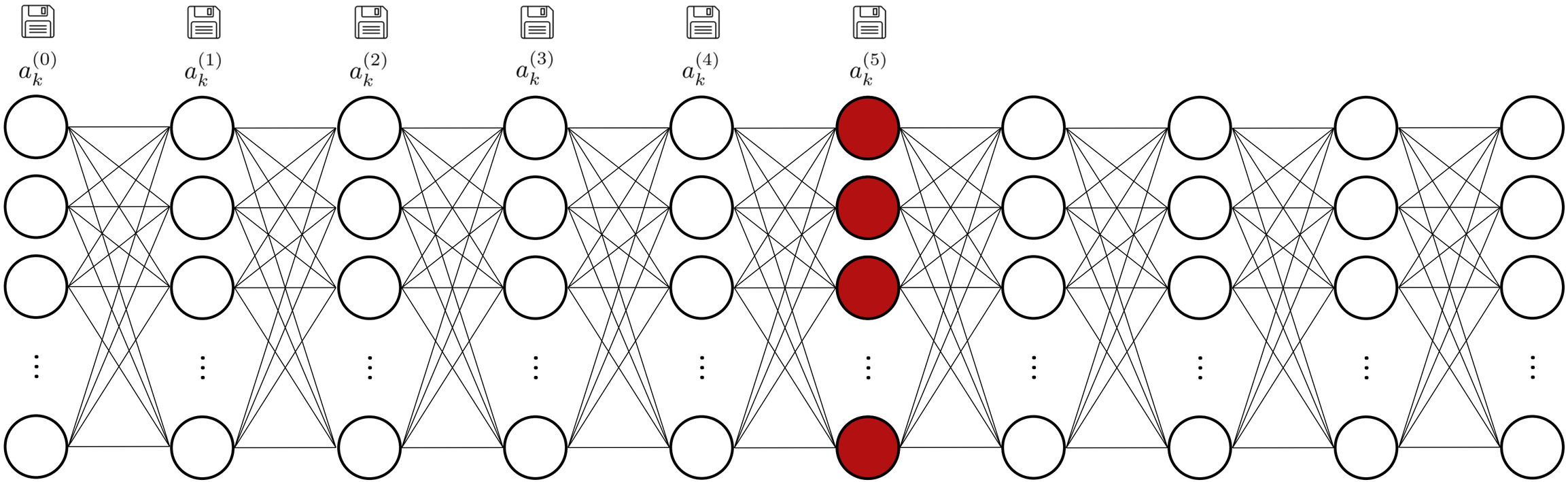


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

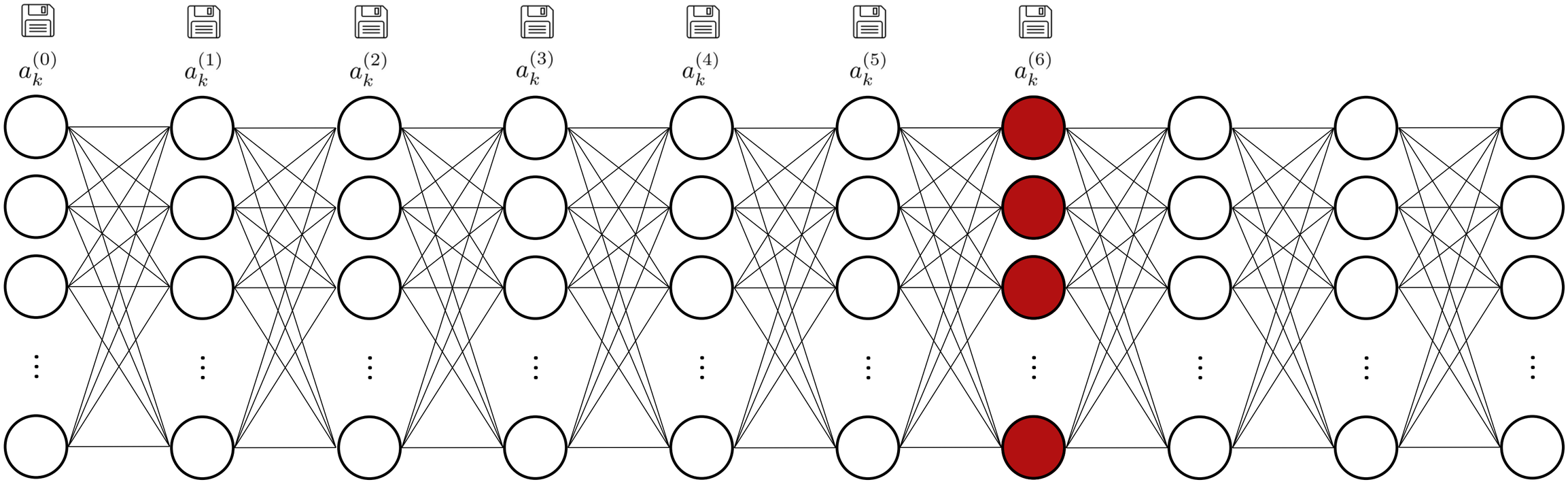


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

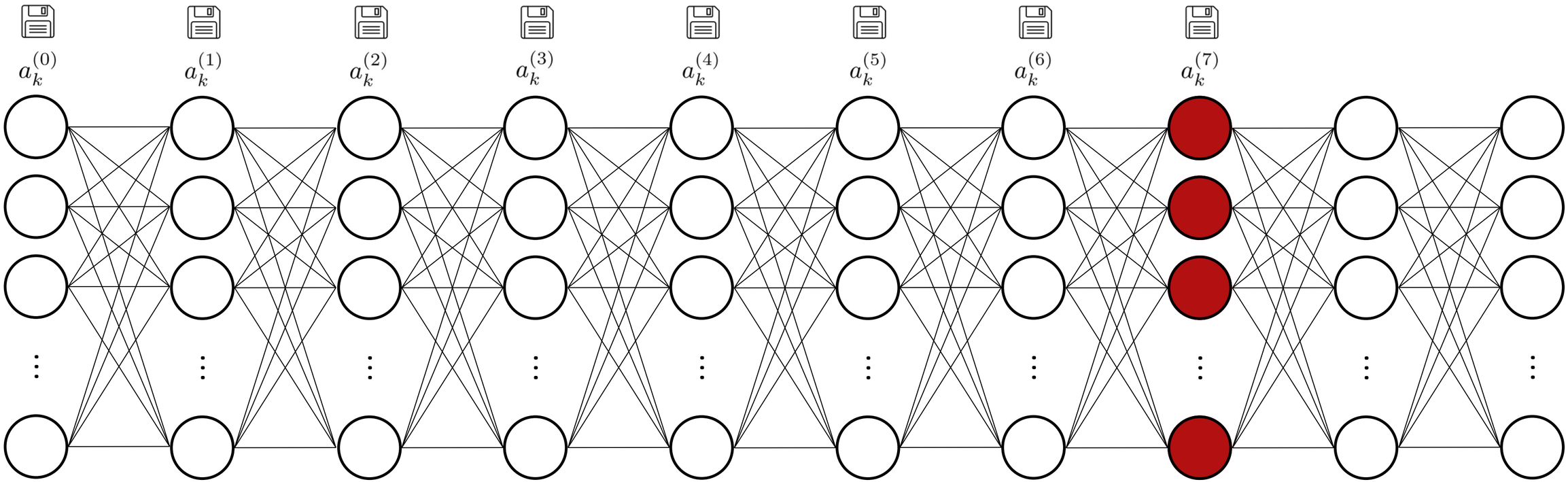


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

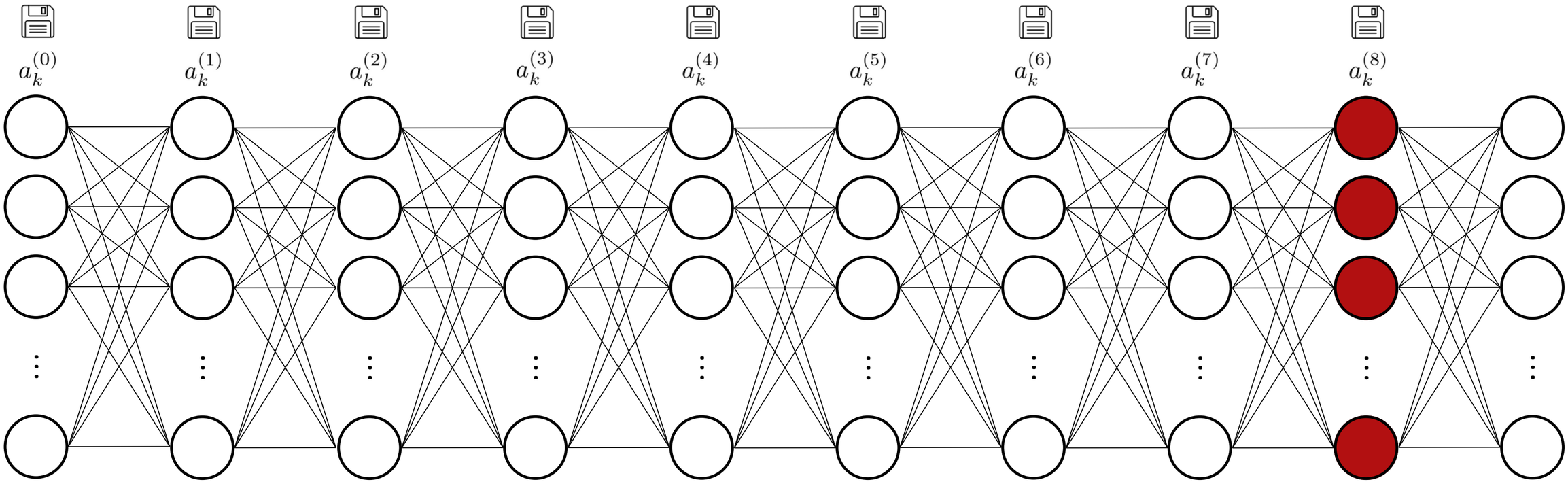


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

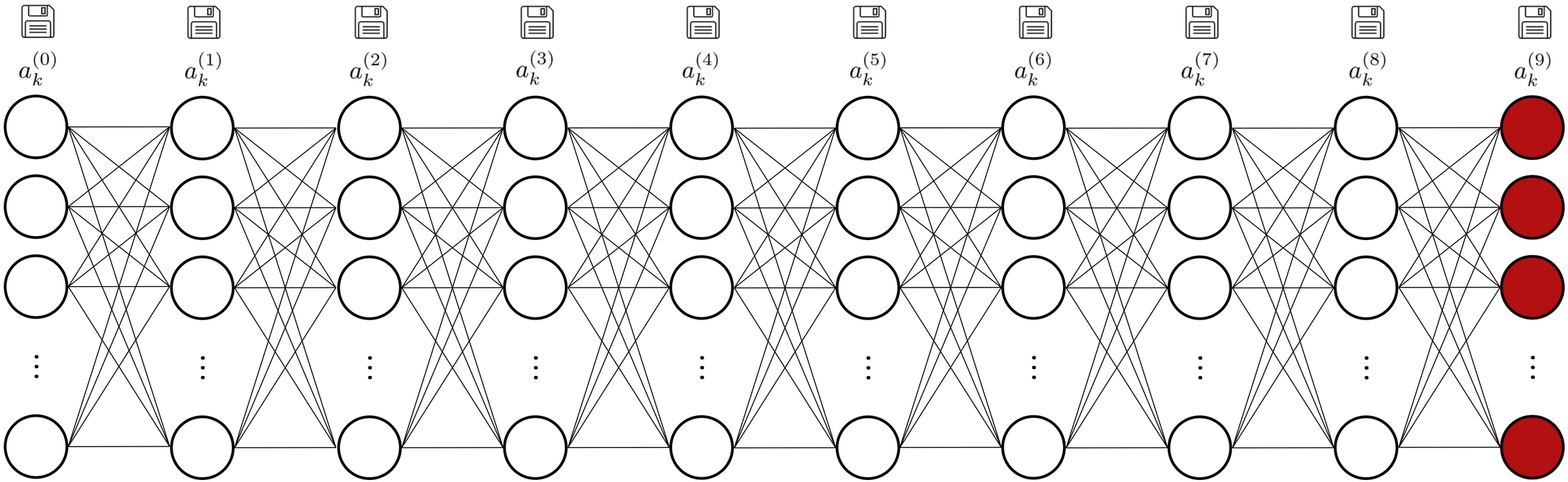


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

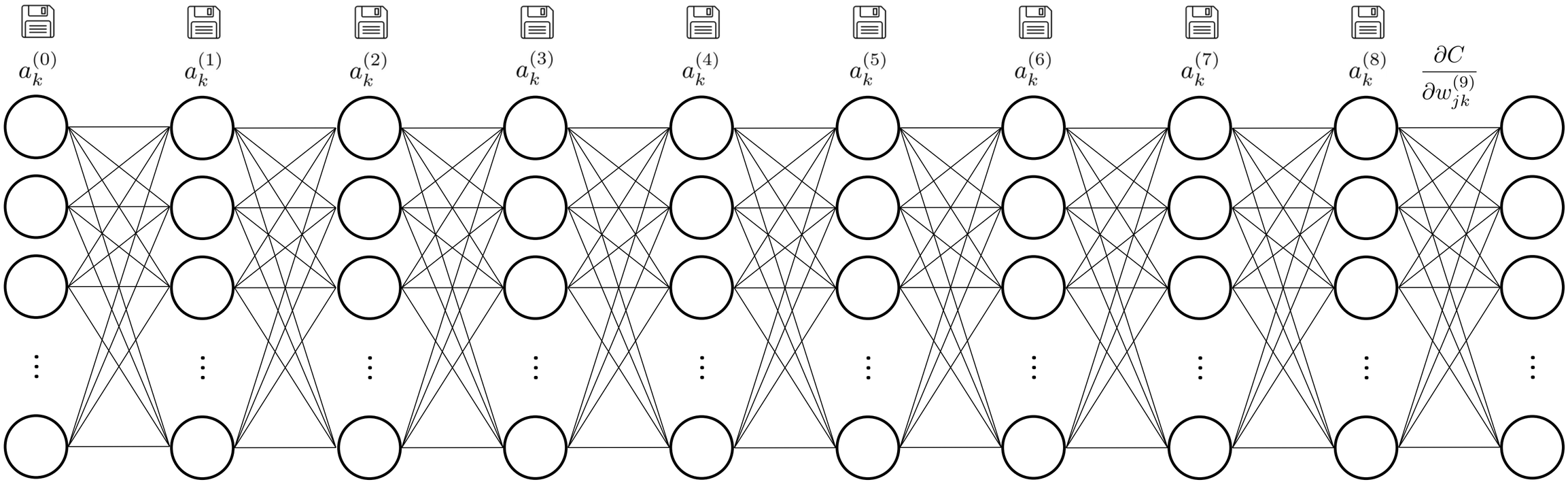


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

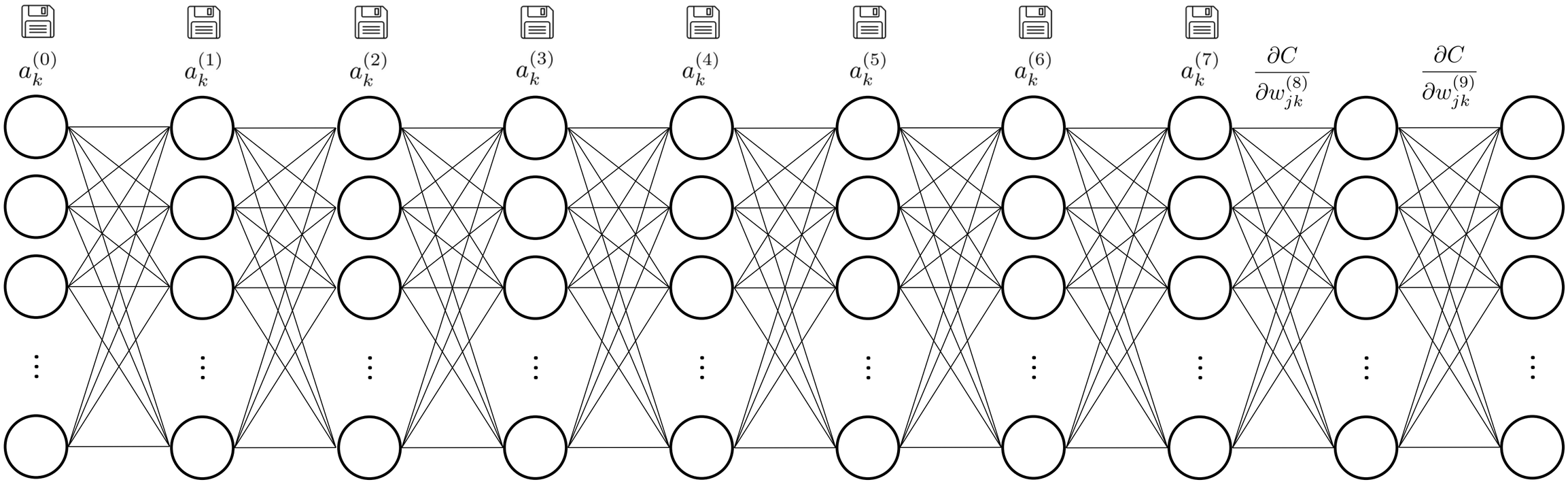


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

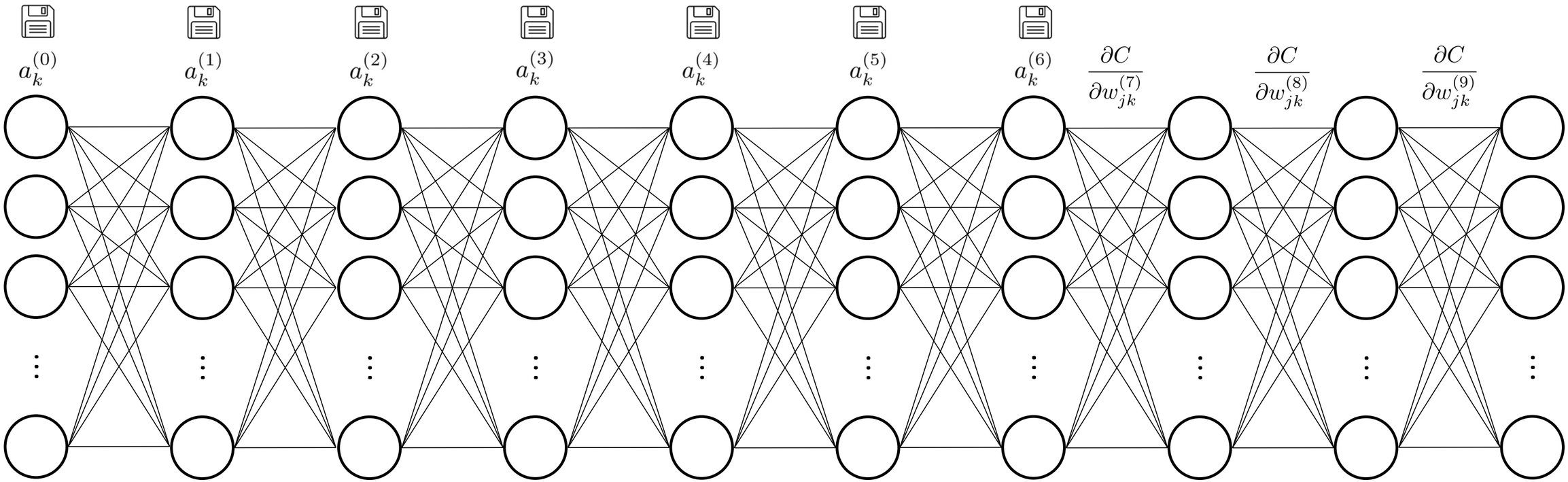


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

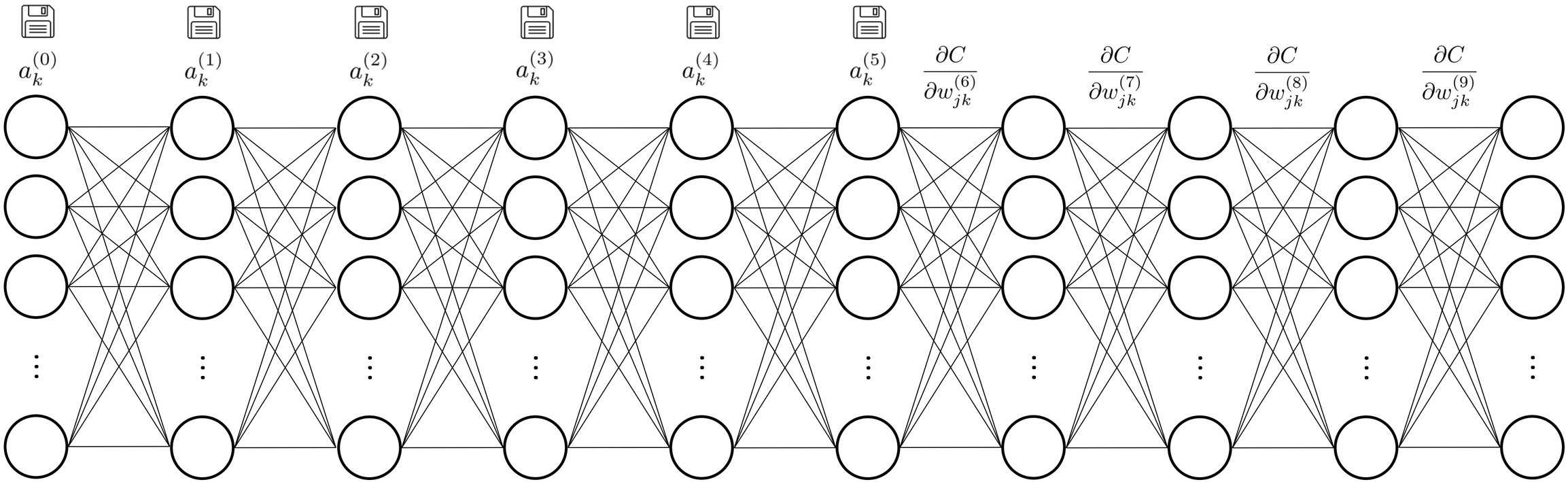


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

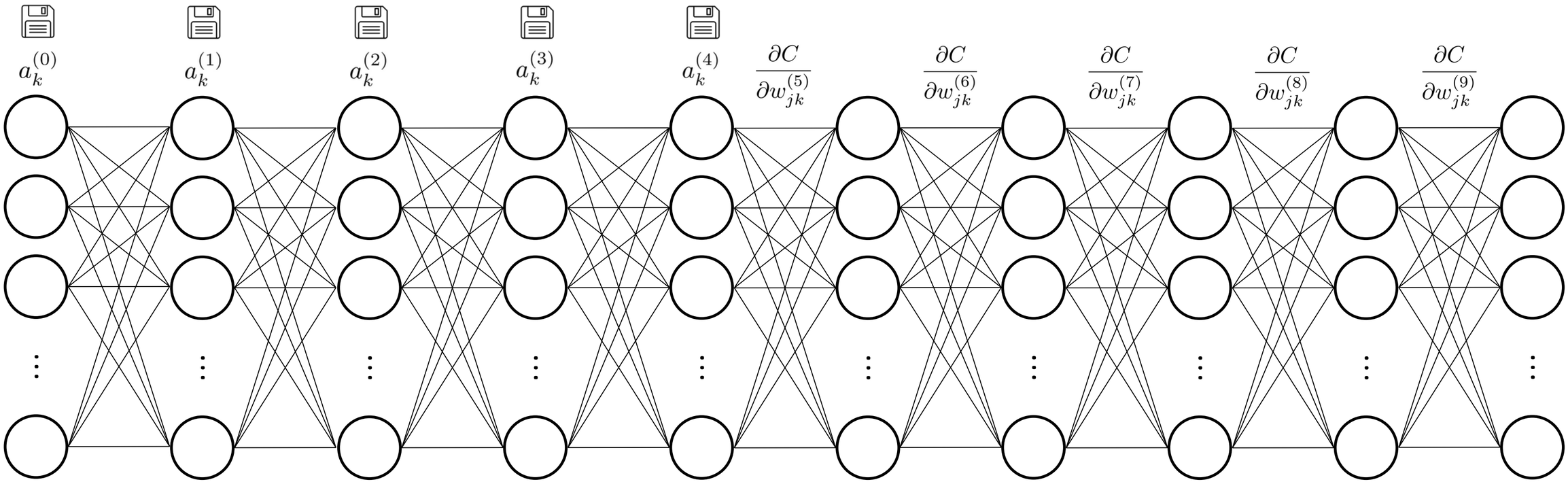


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

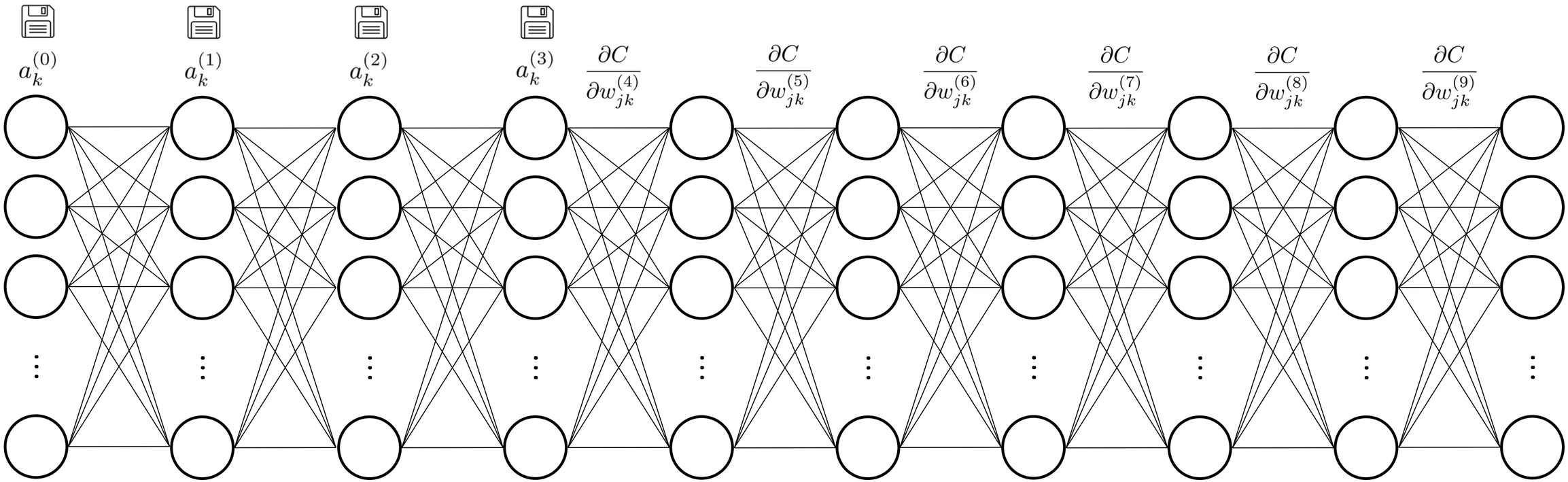


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

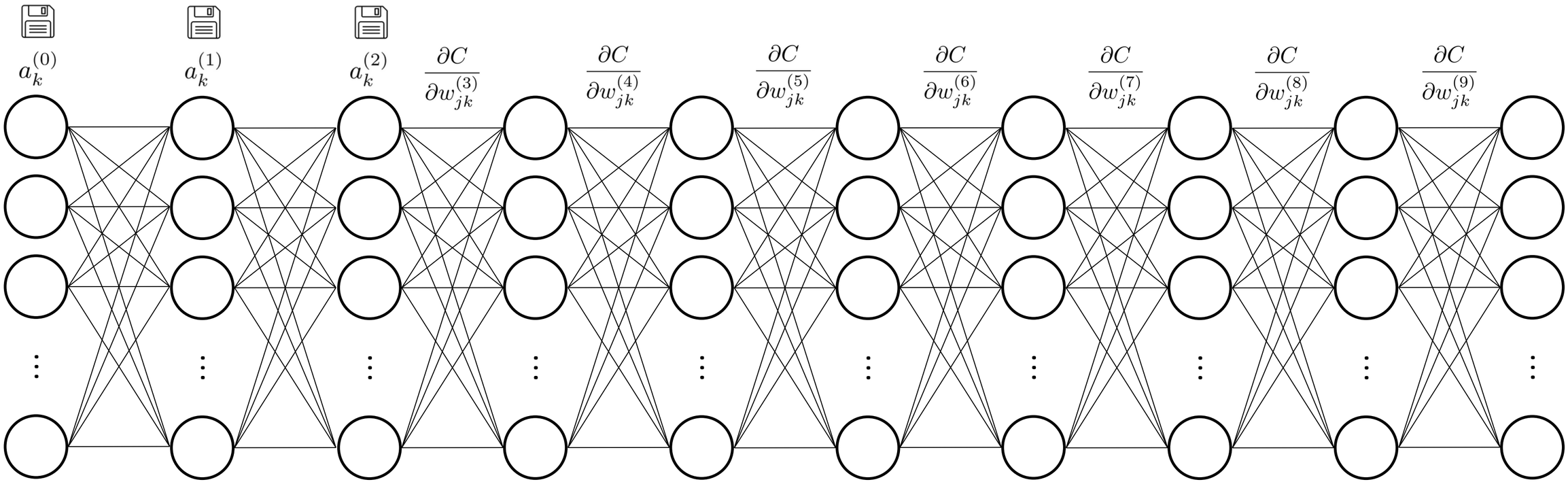


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

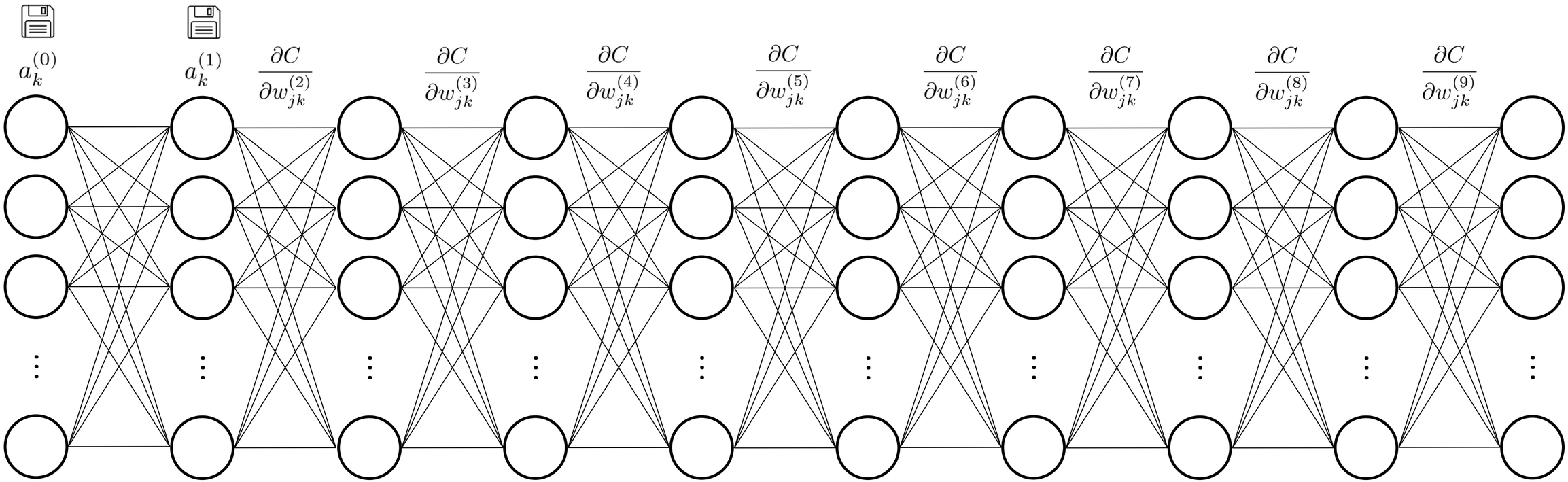


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

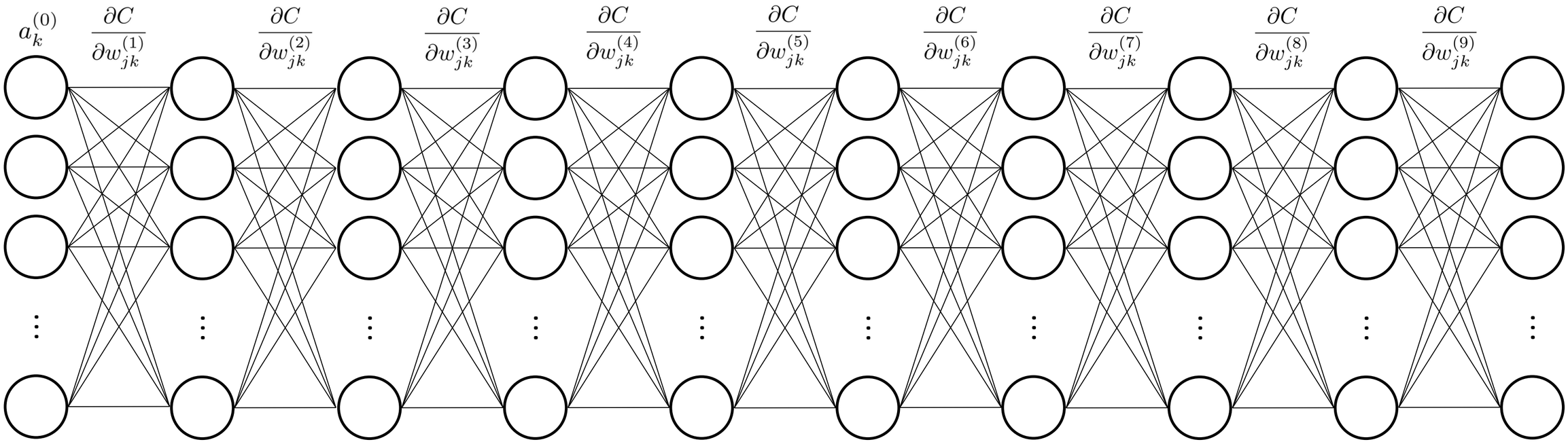


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

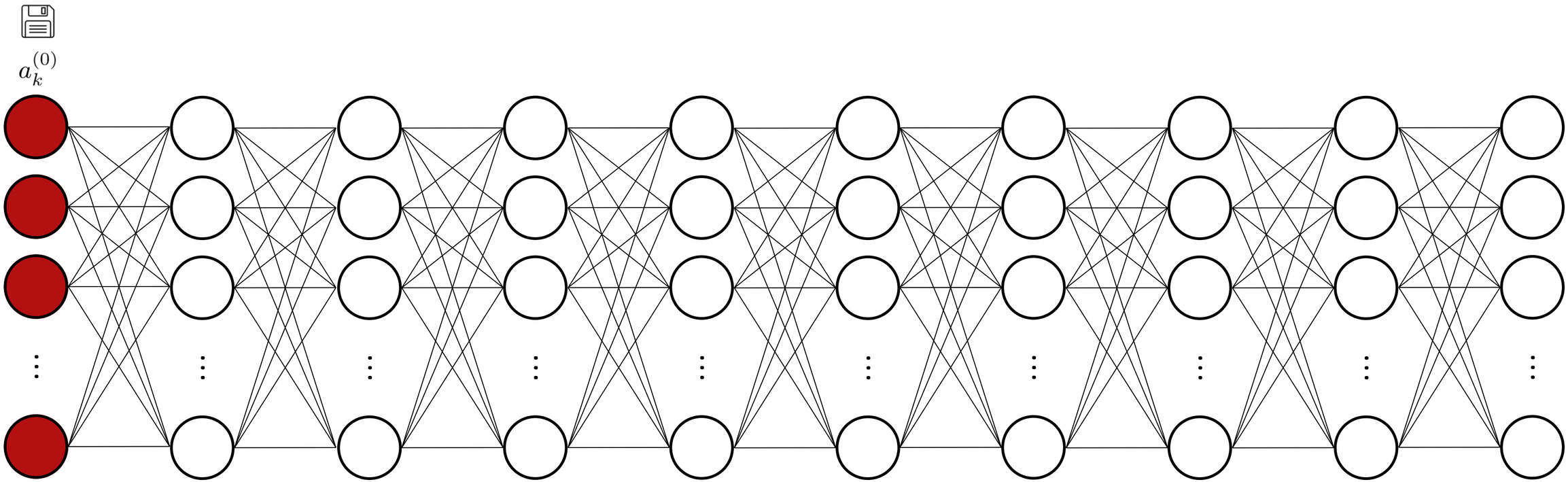


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

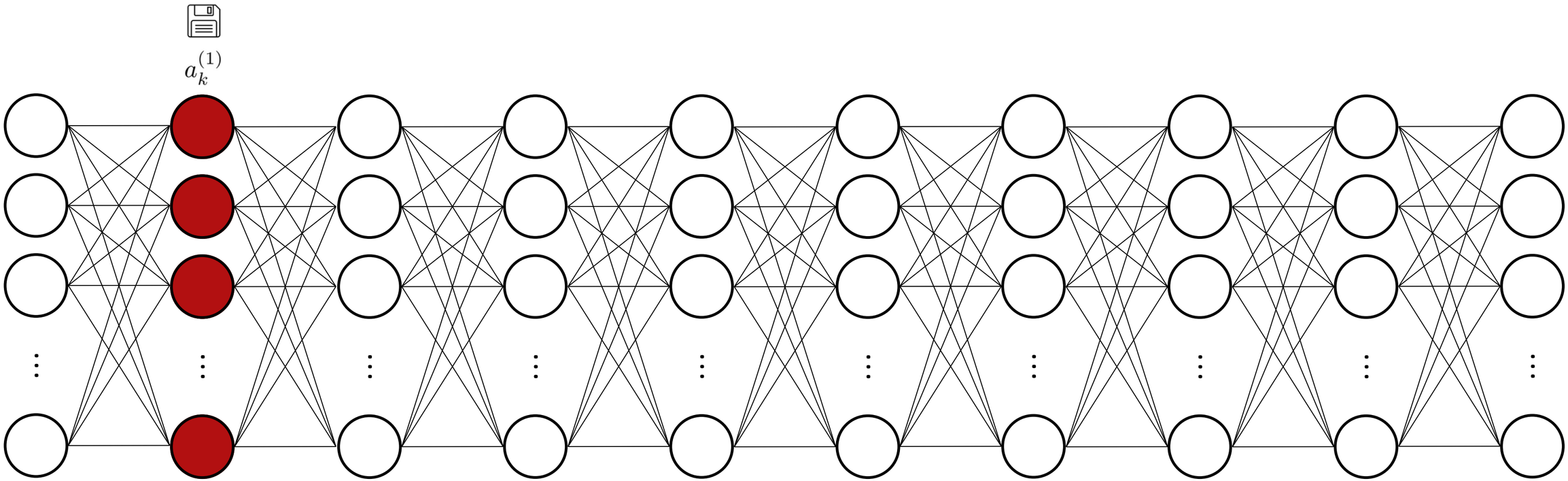


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} \, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)} \, w_{kj}^{(l+1)} \, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2 \left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

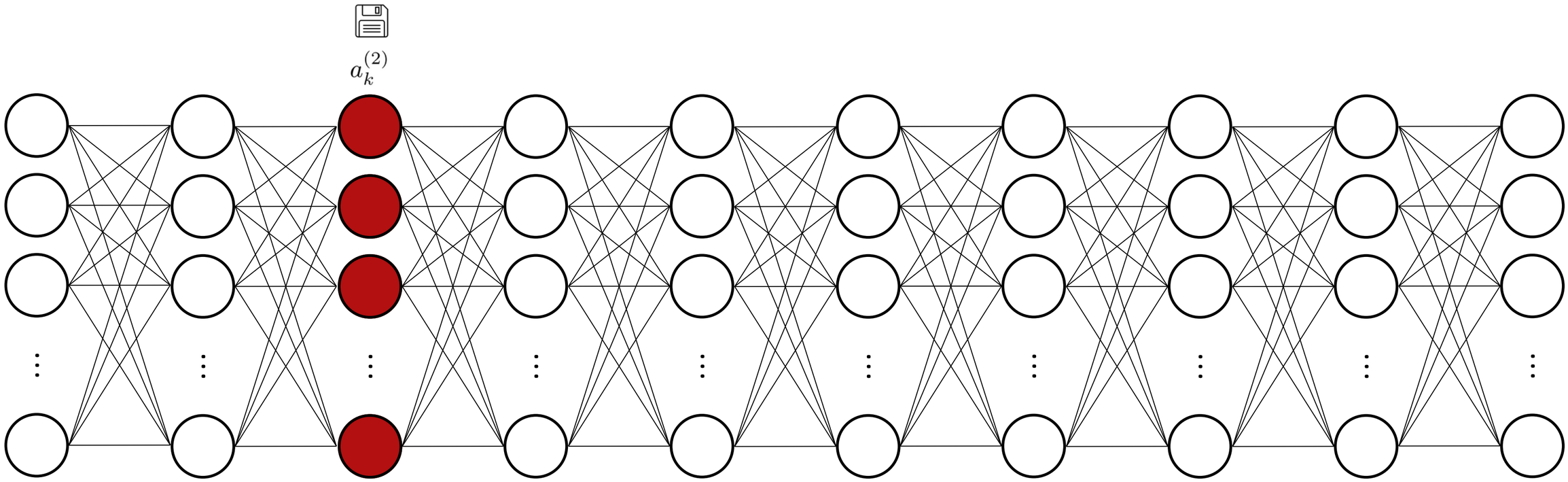


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

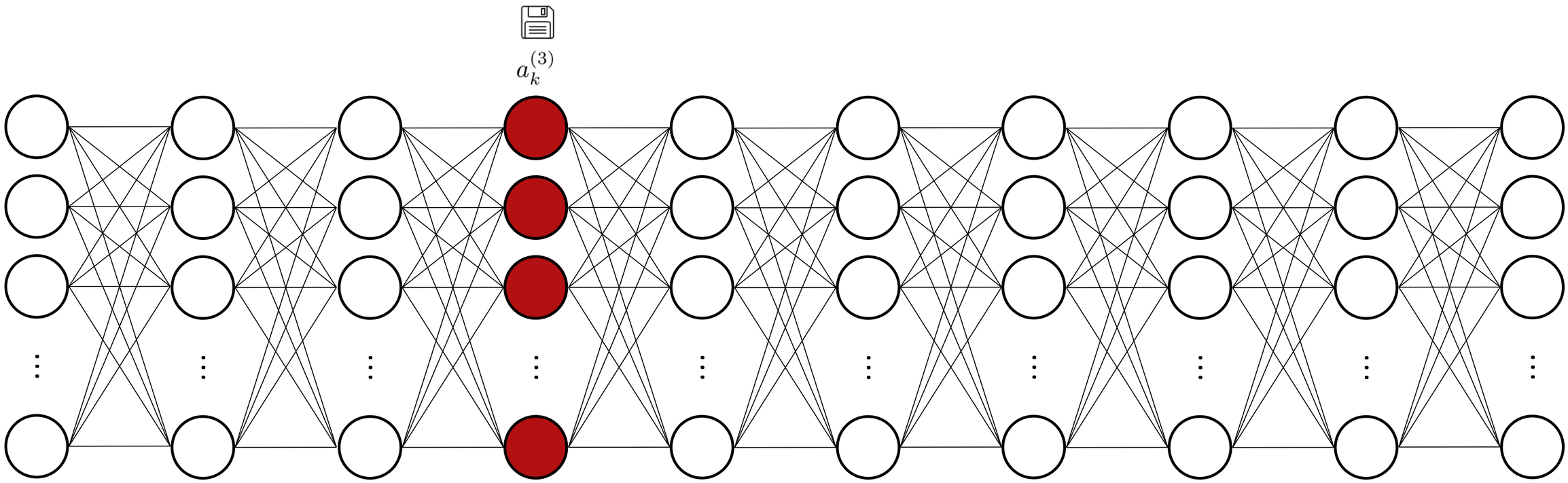


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

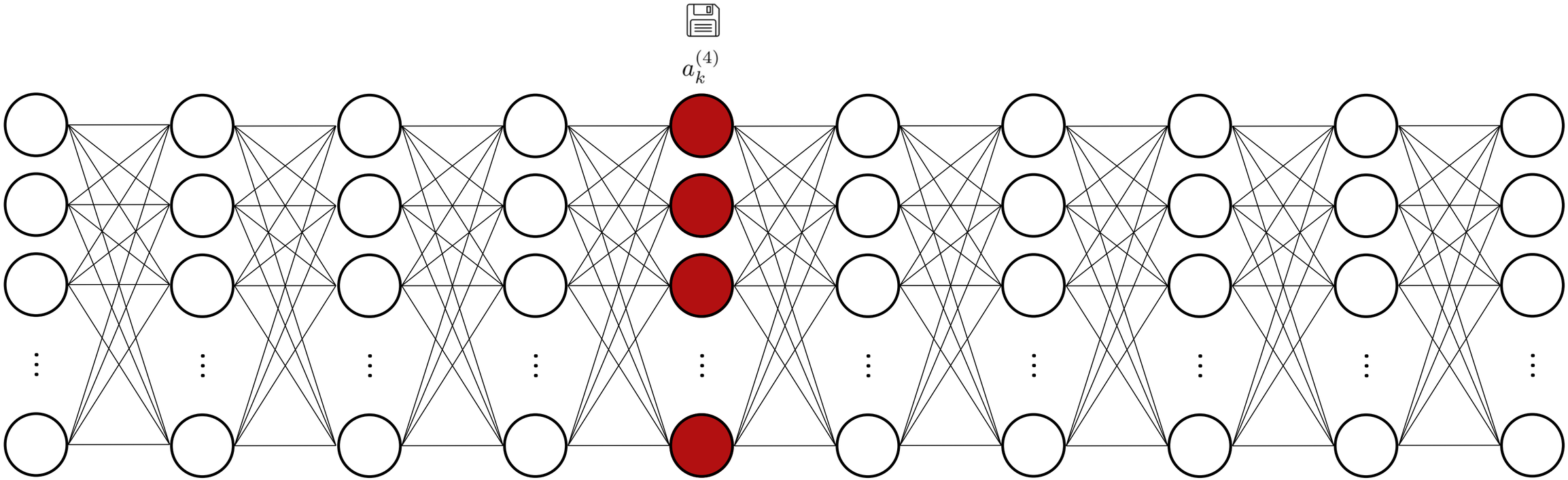


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

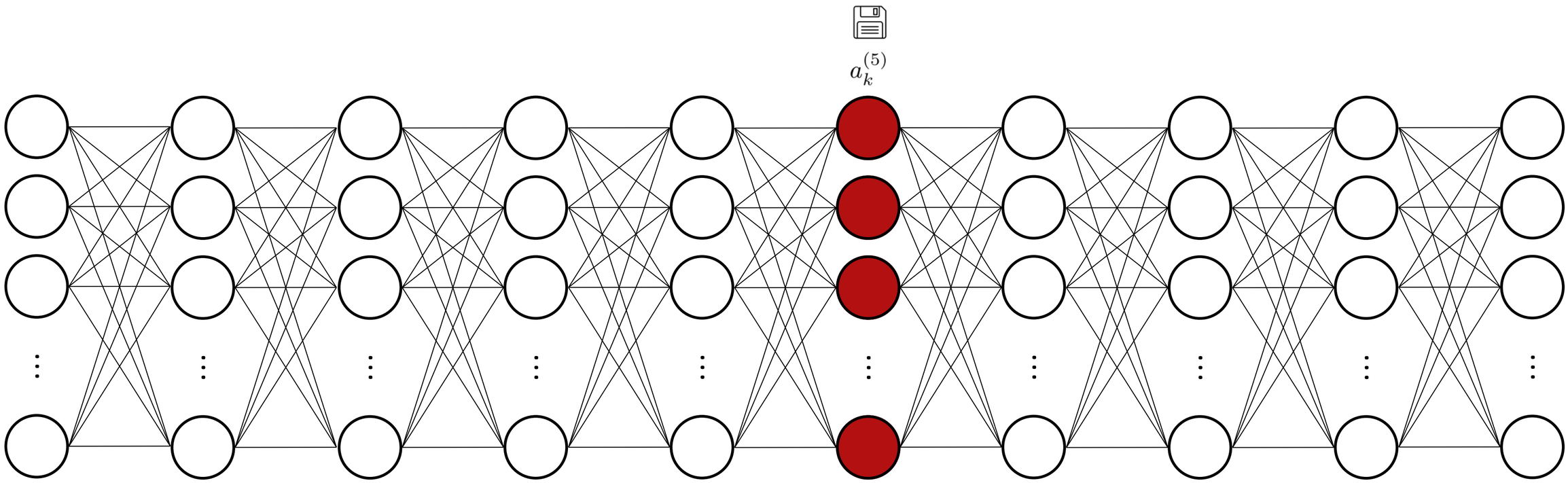


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

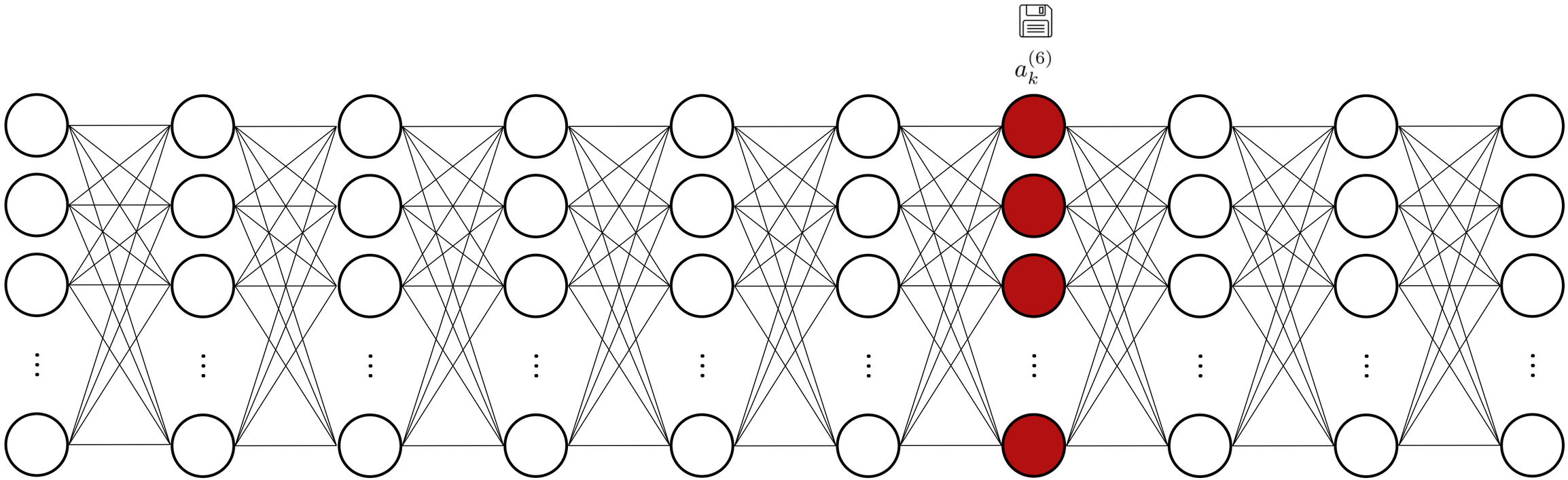


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

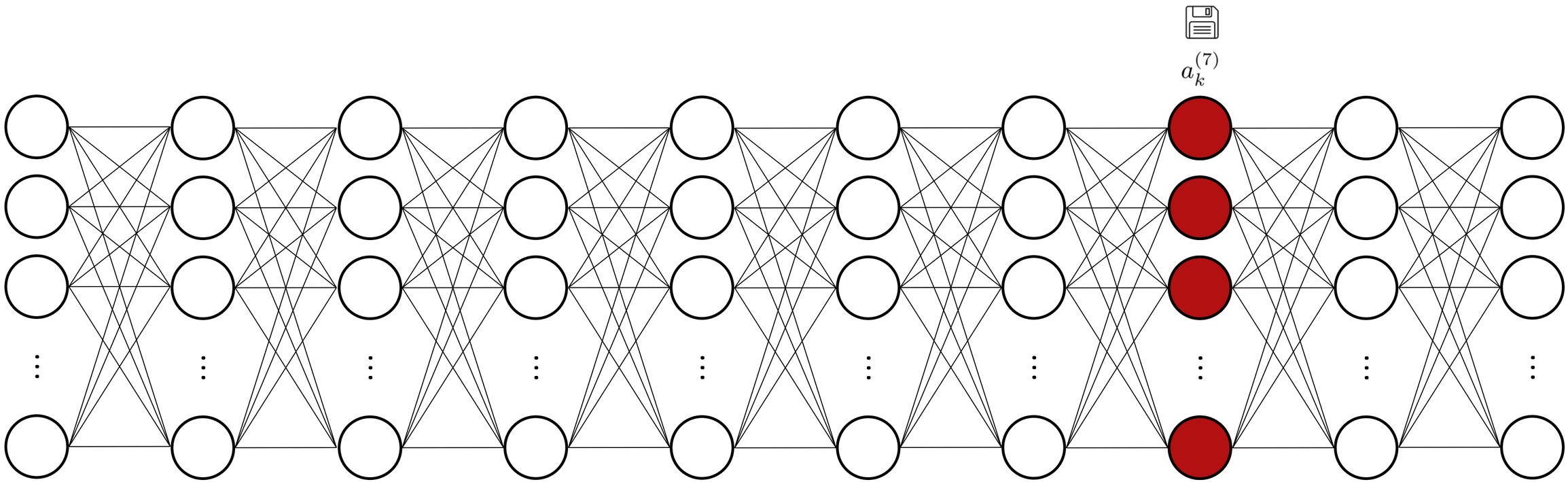


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

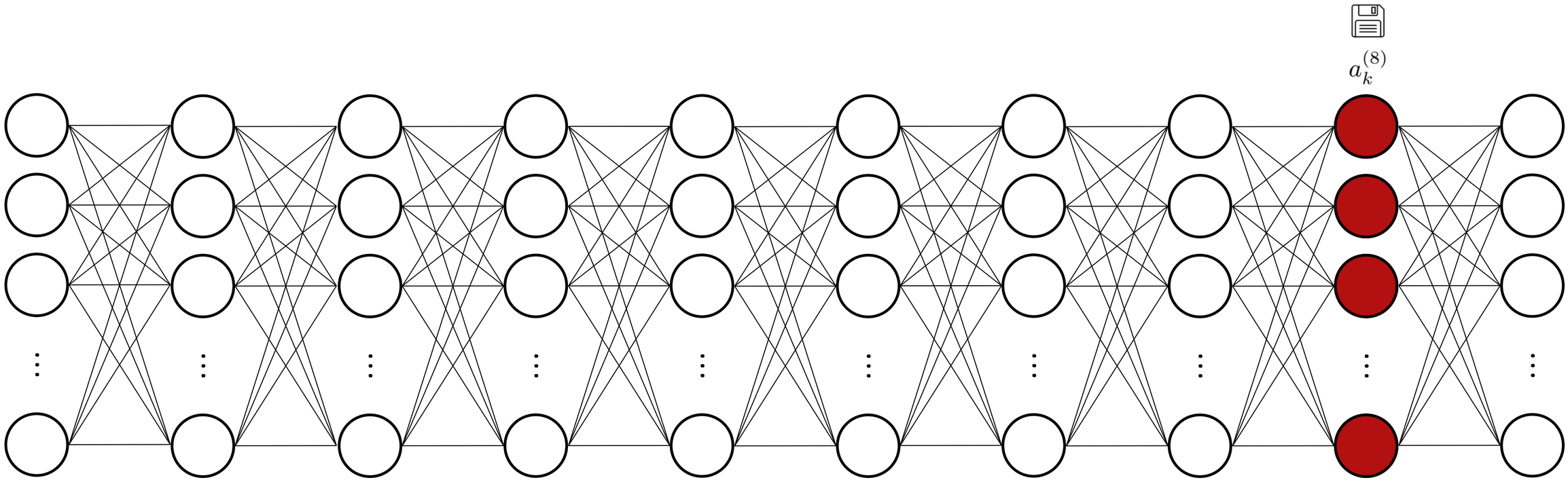


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

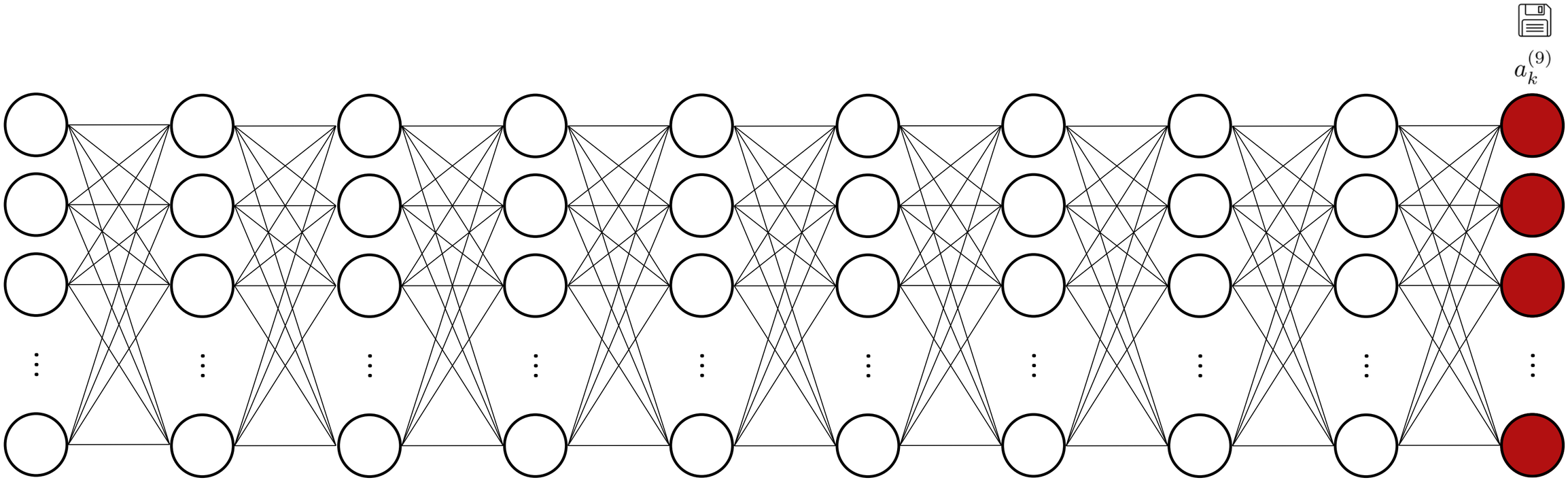


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

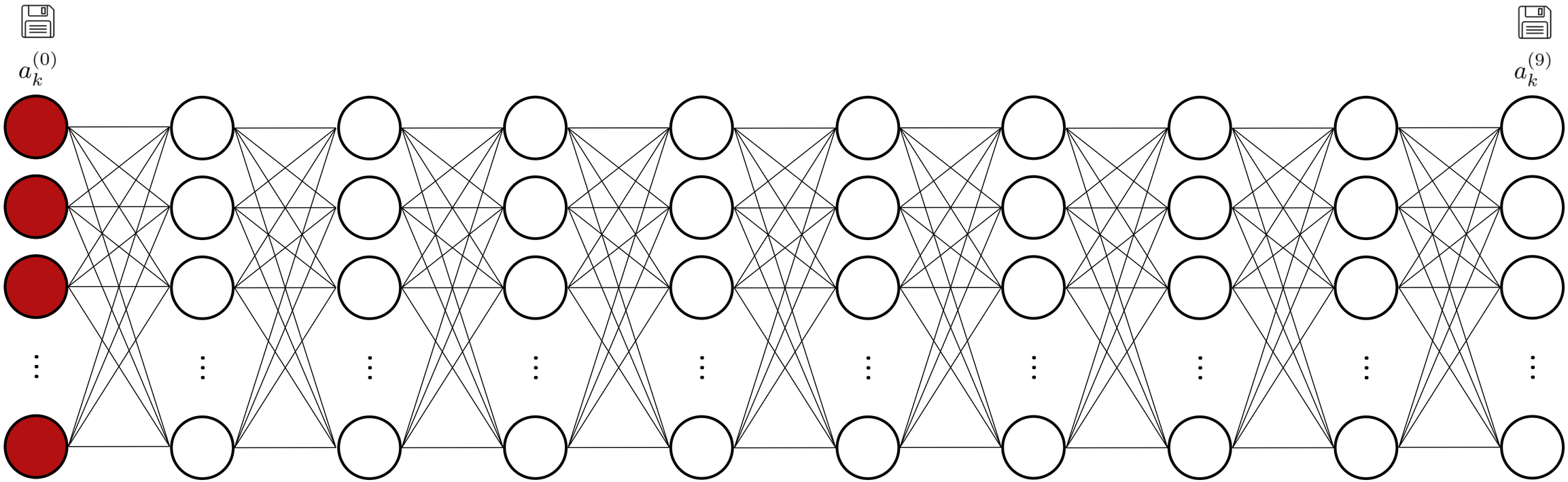


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

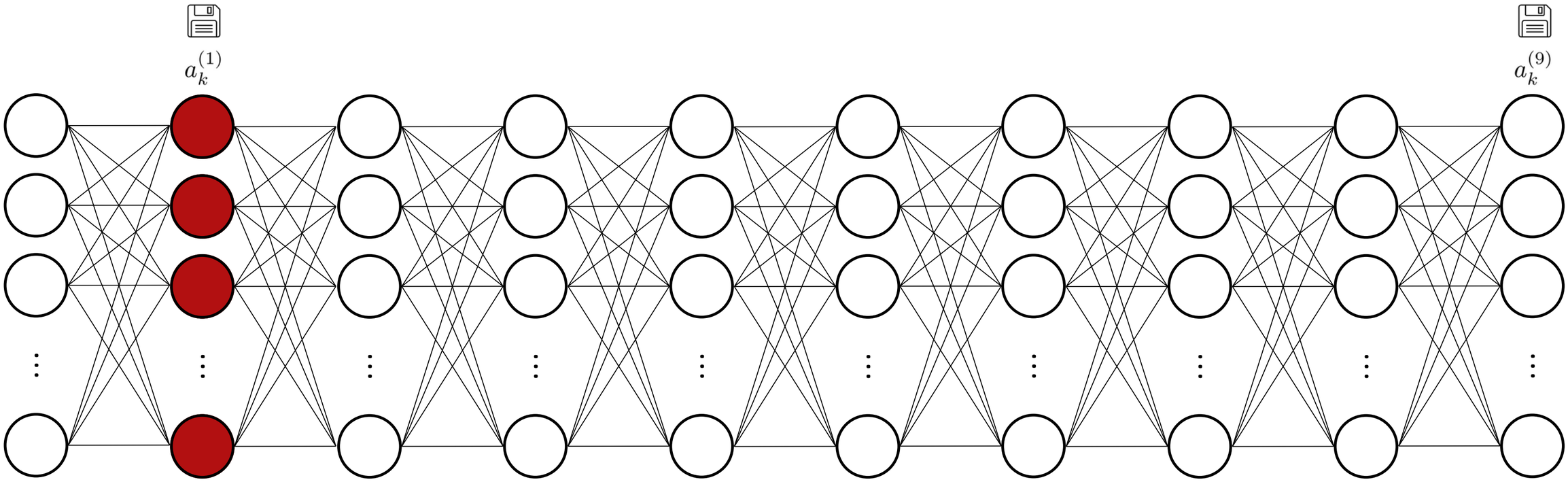


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

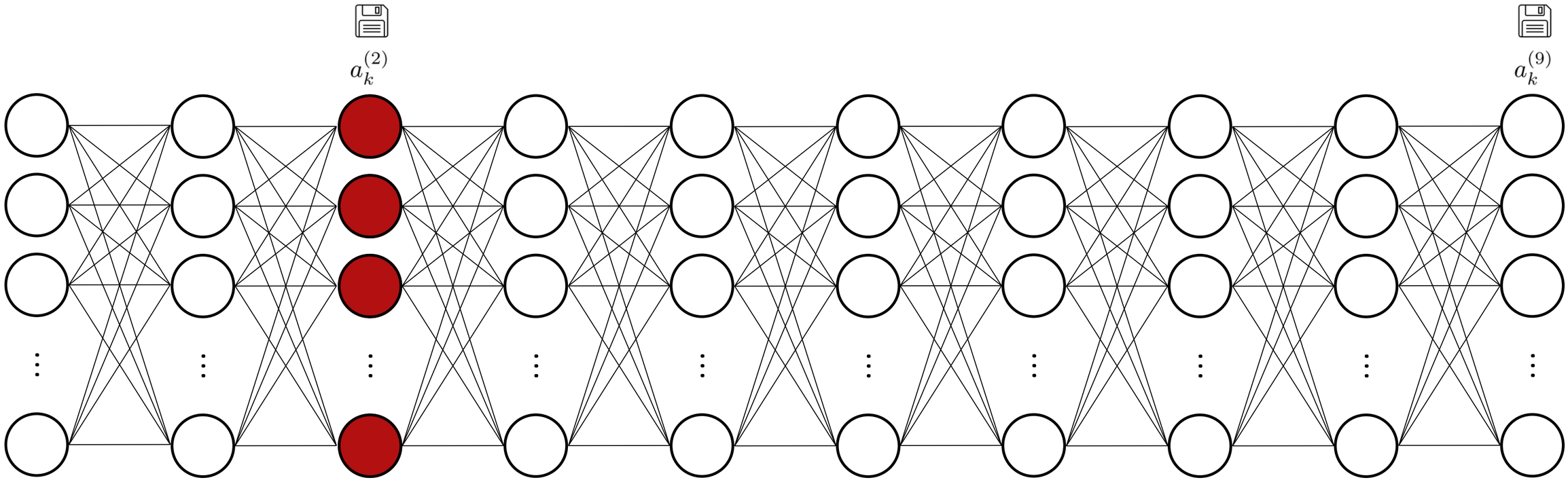


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

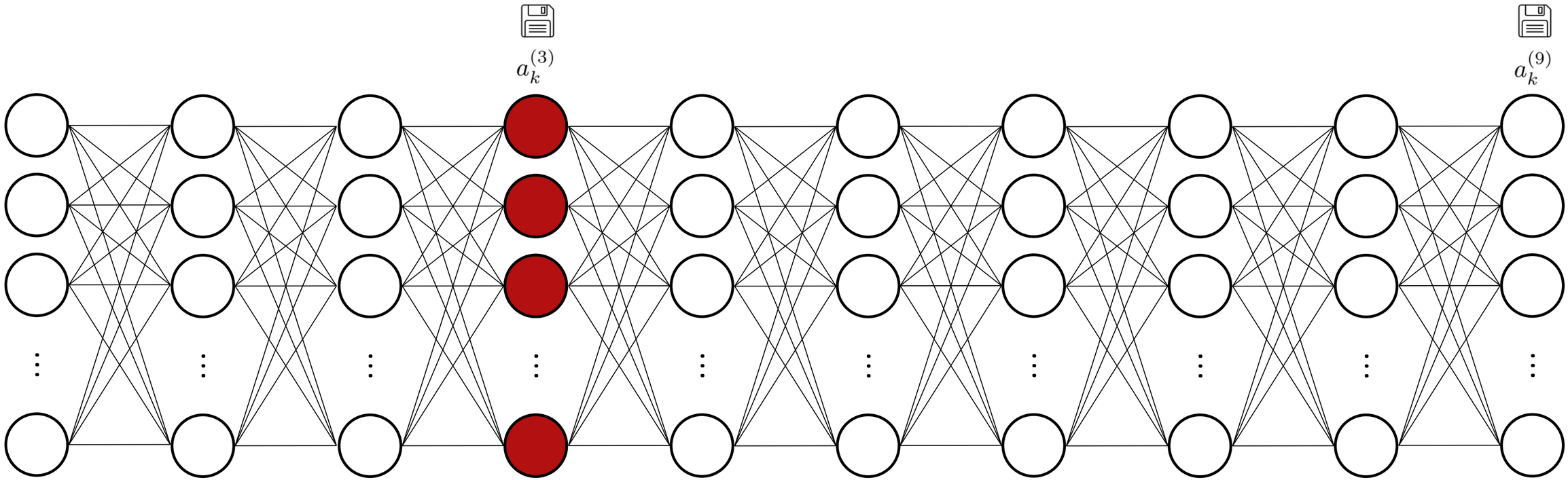


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

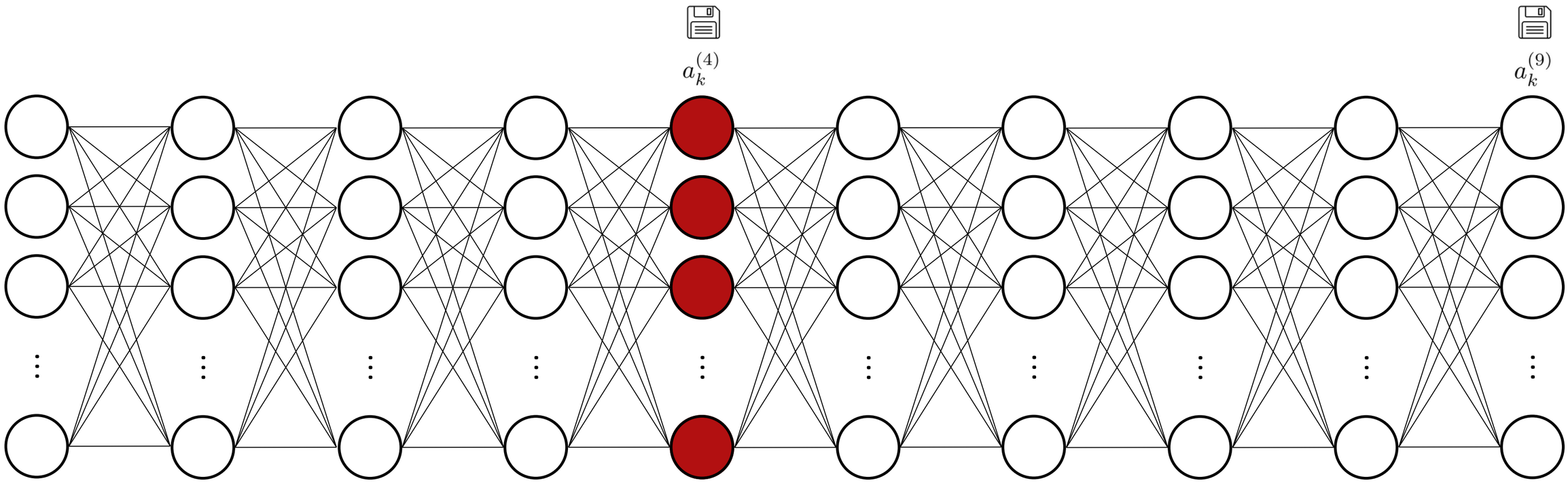


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

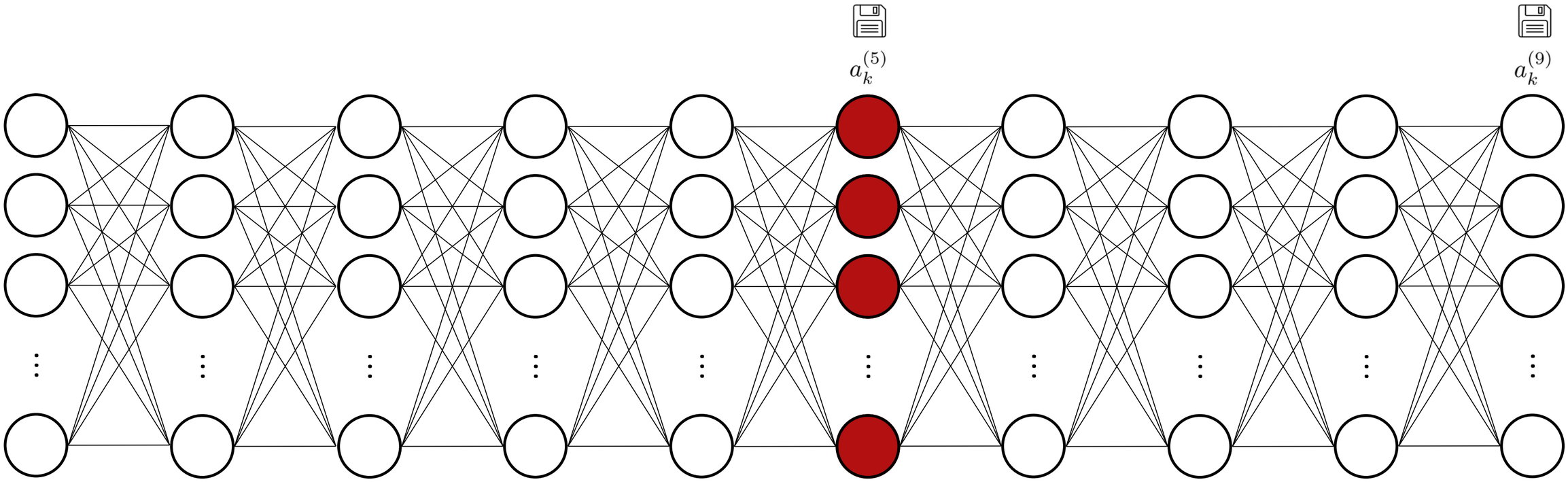


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

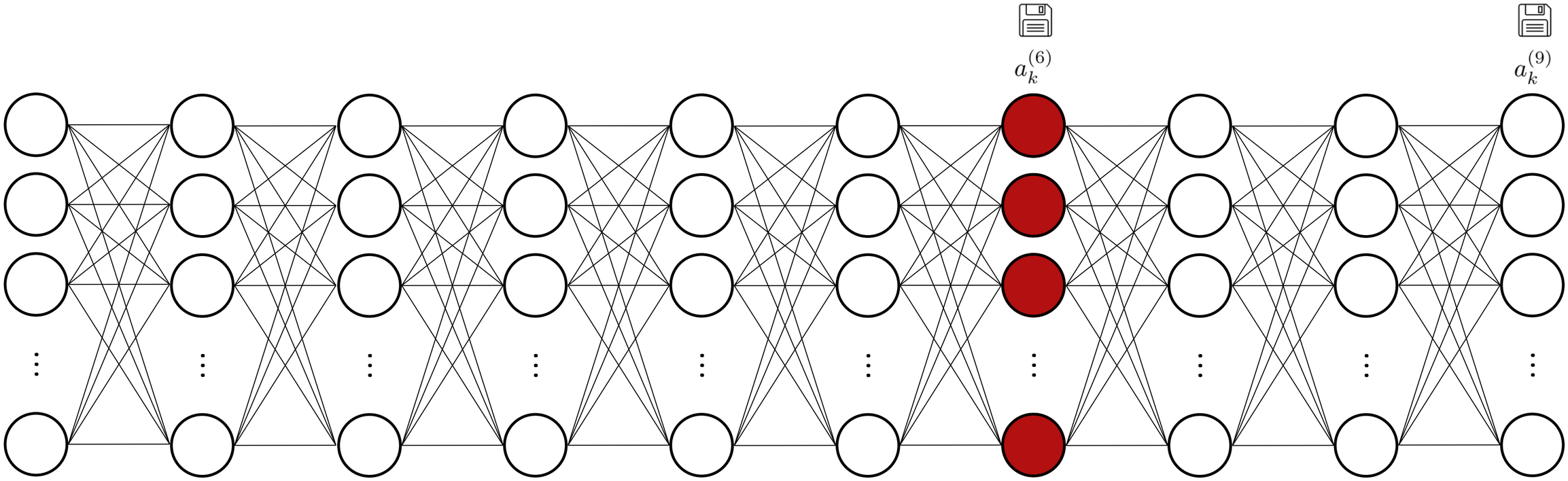


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

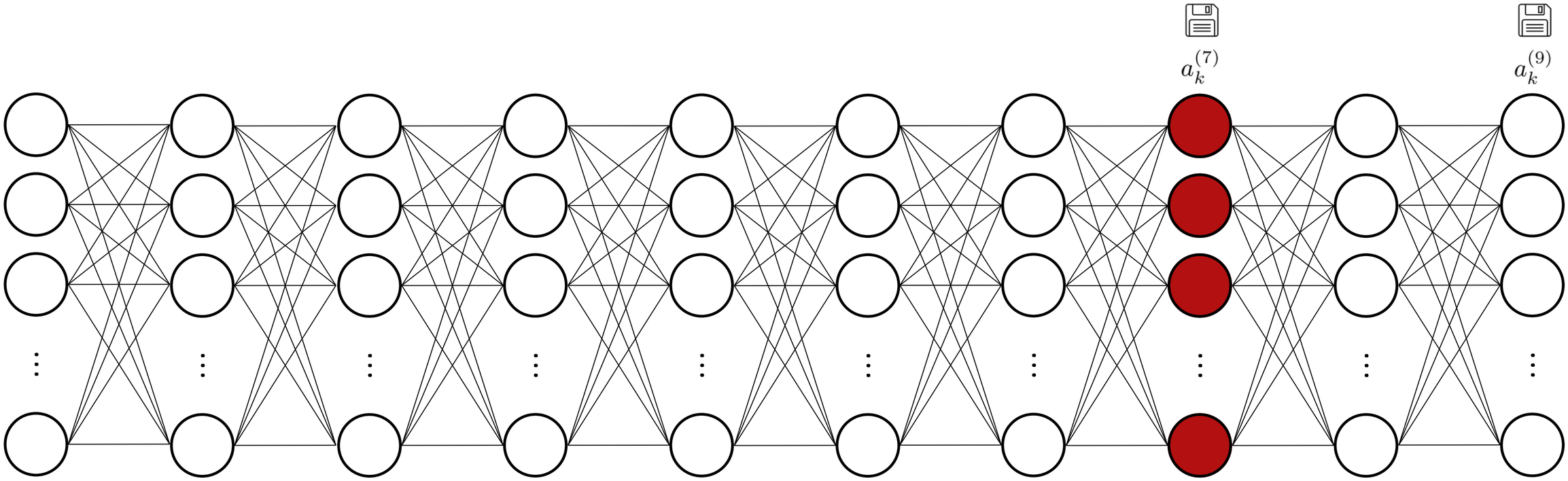


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

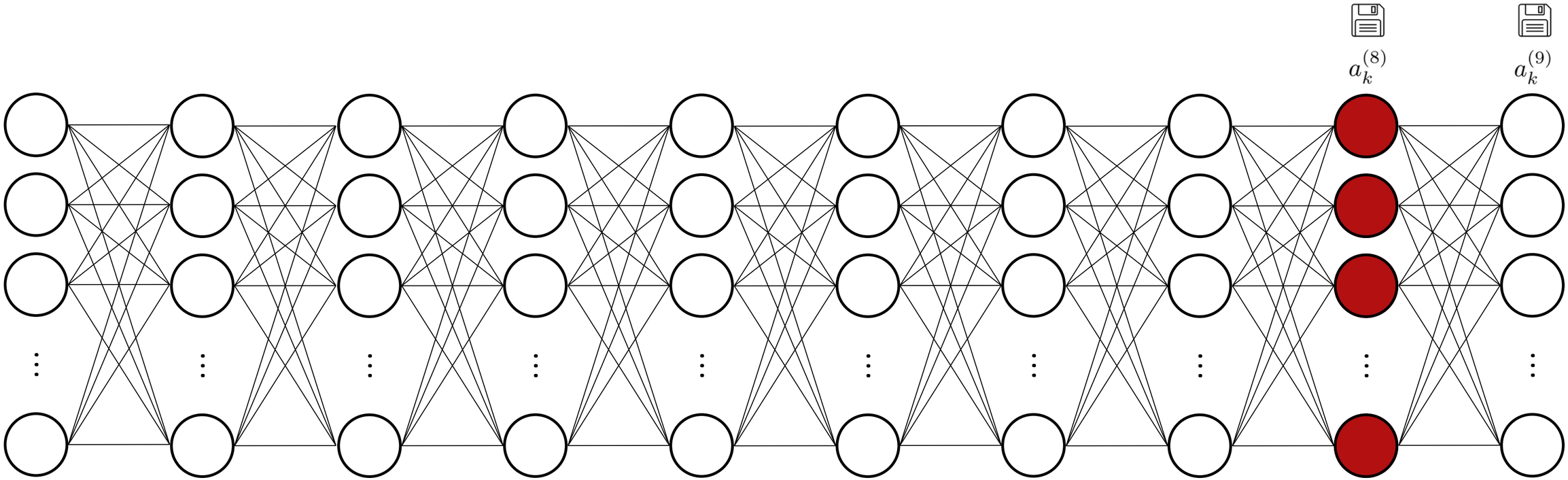


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

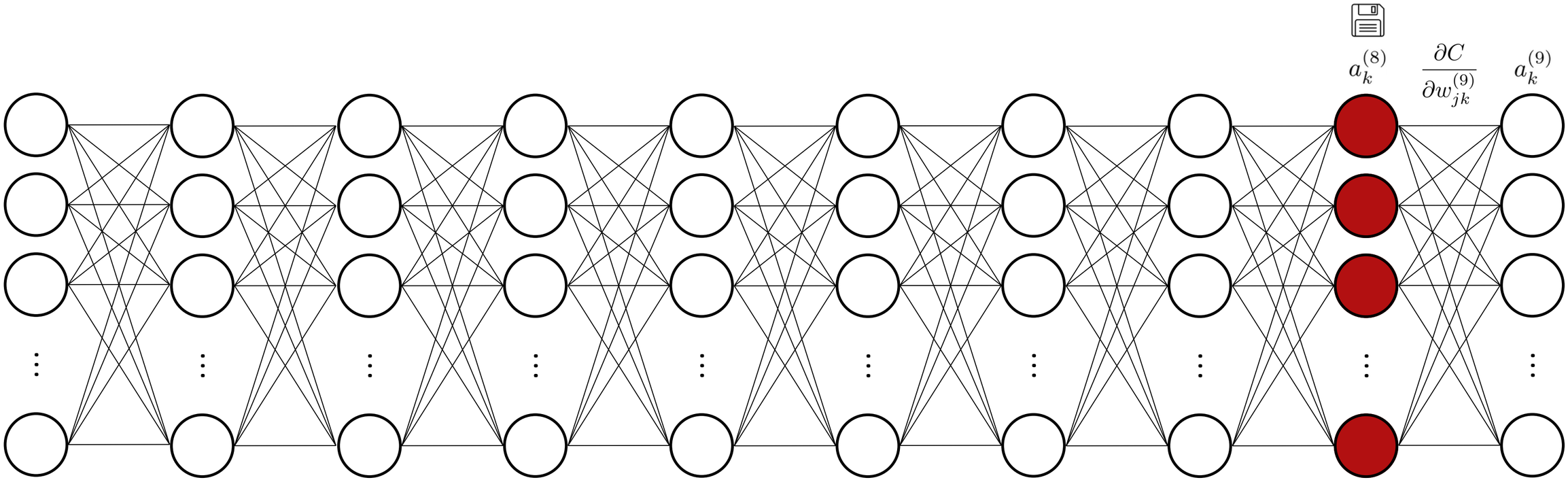


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

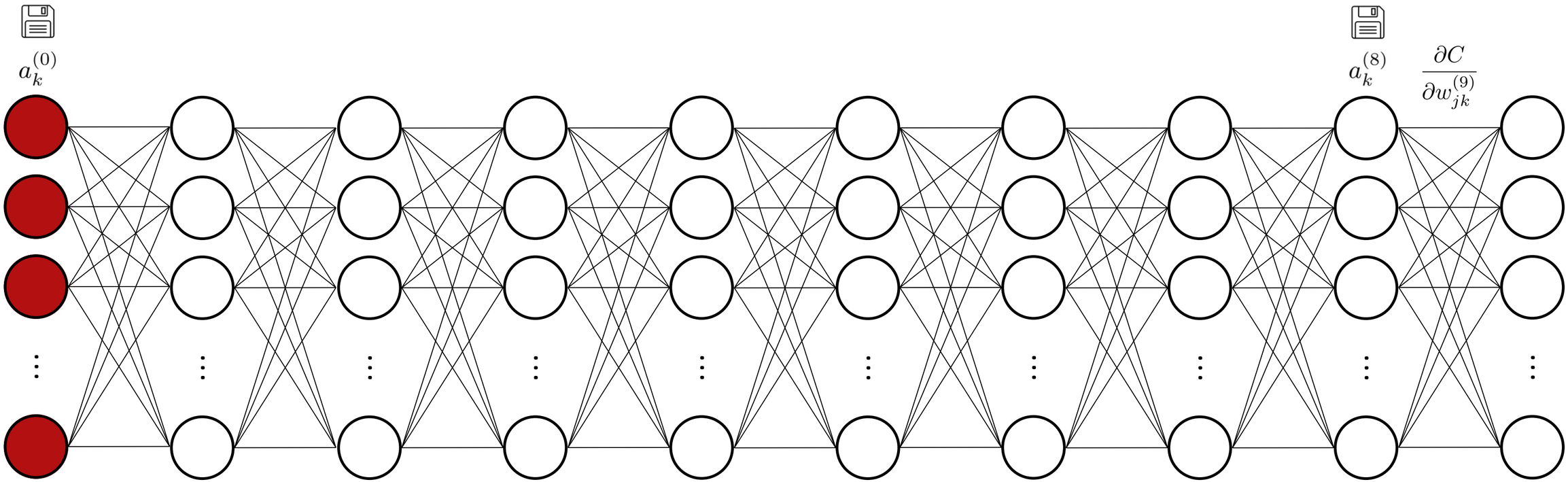


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

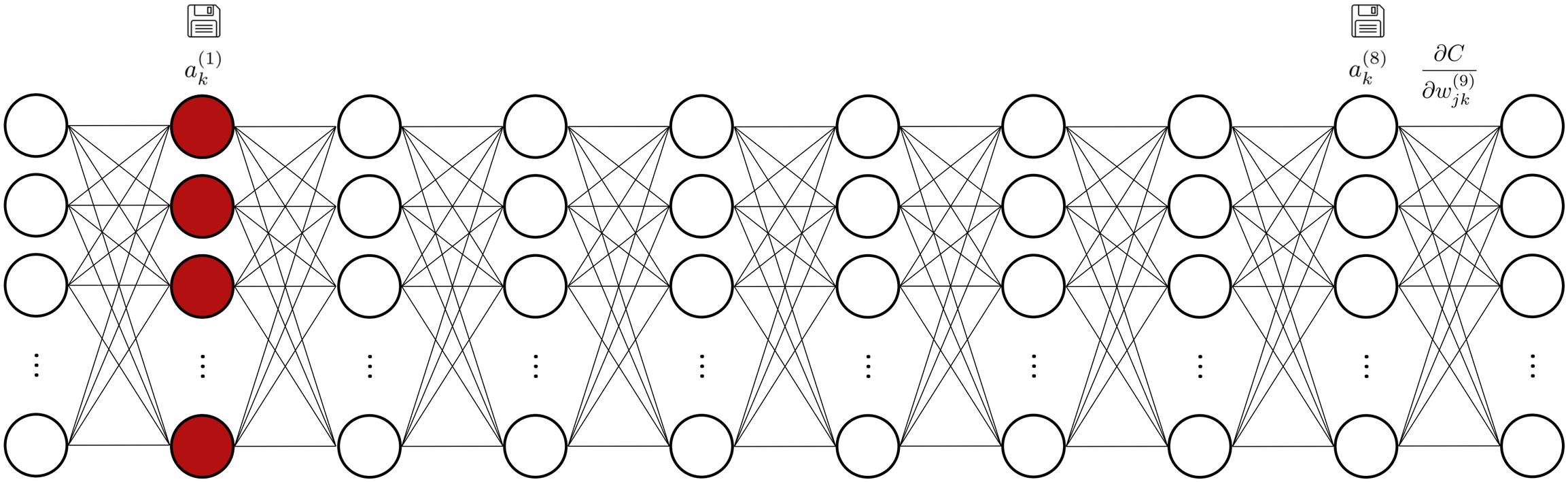


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

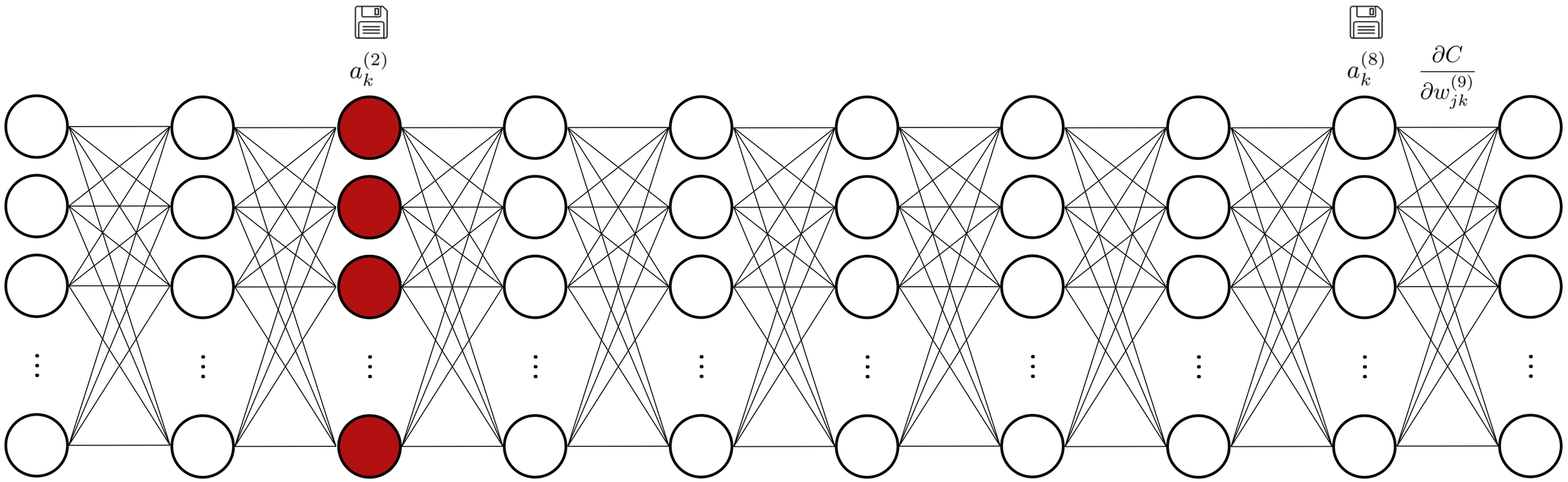


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

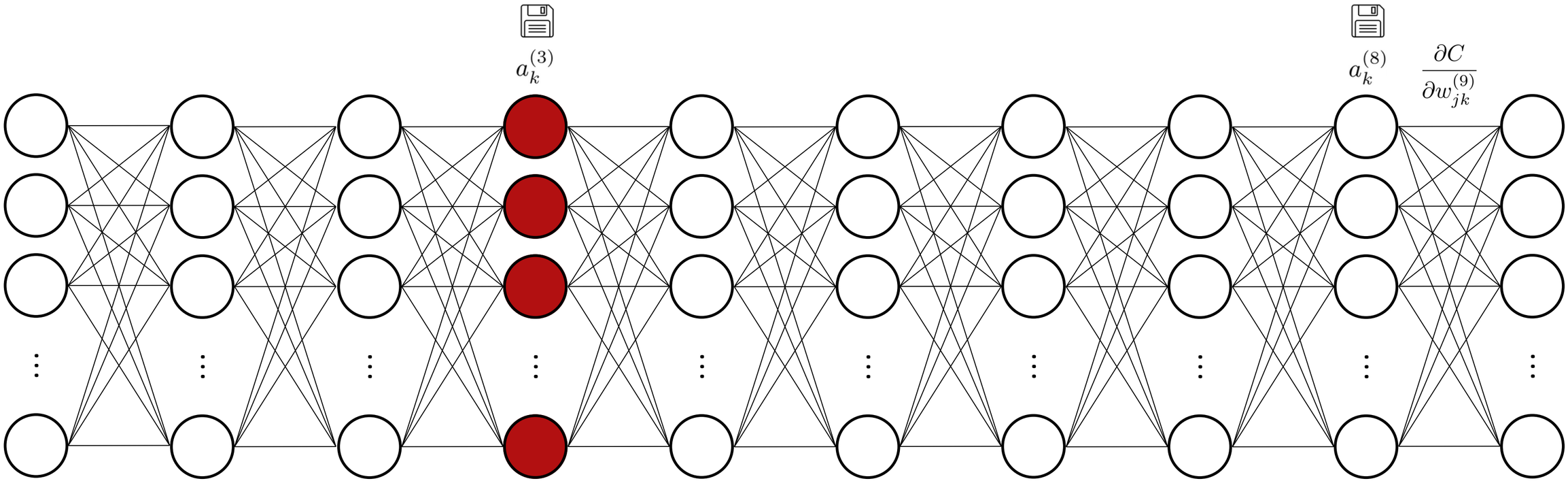


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

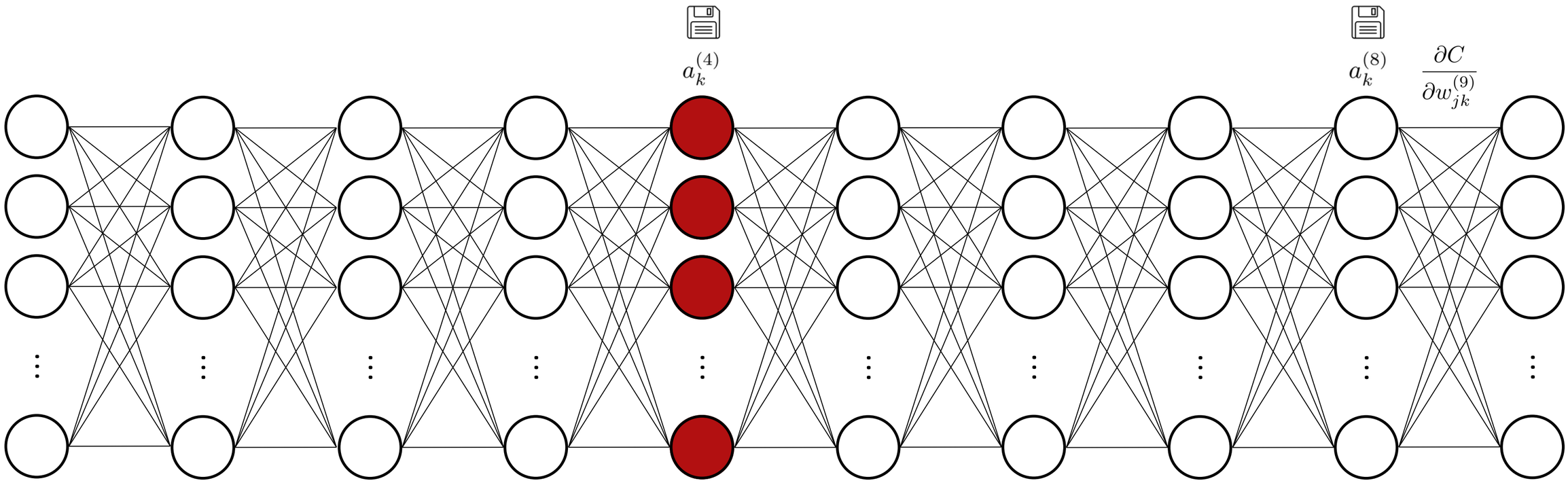


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

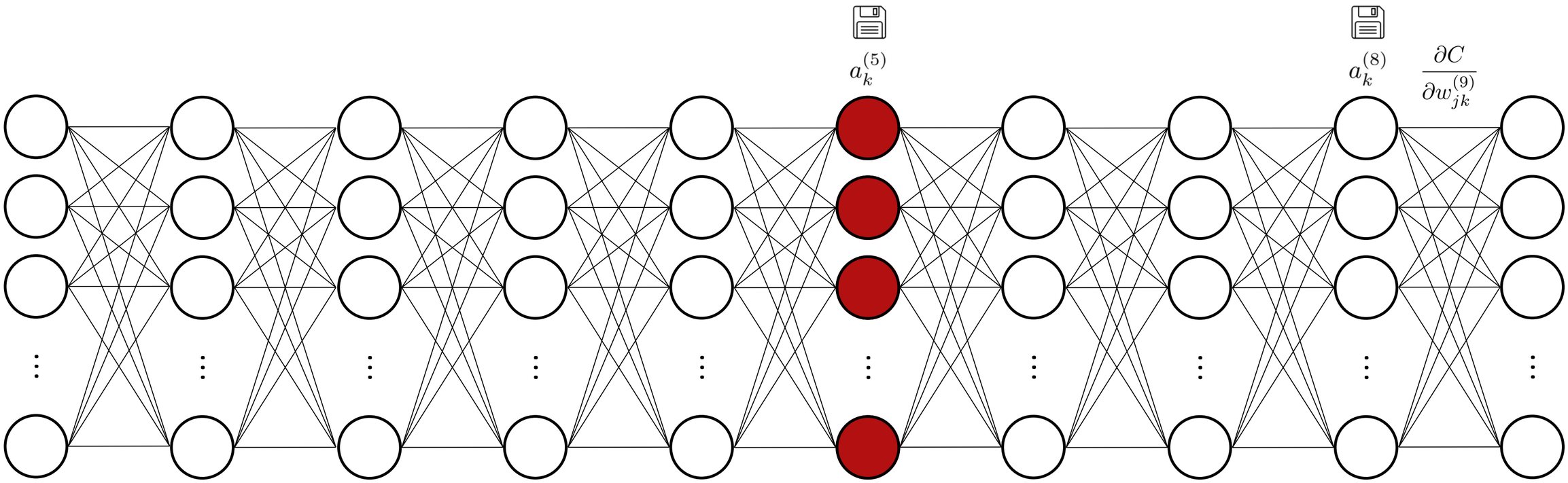


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

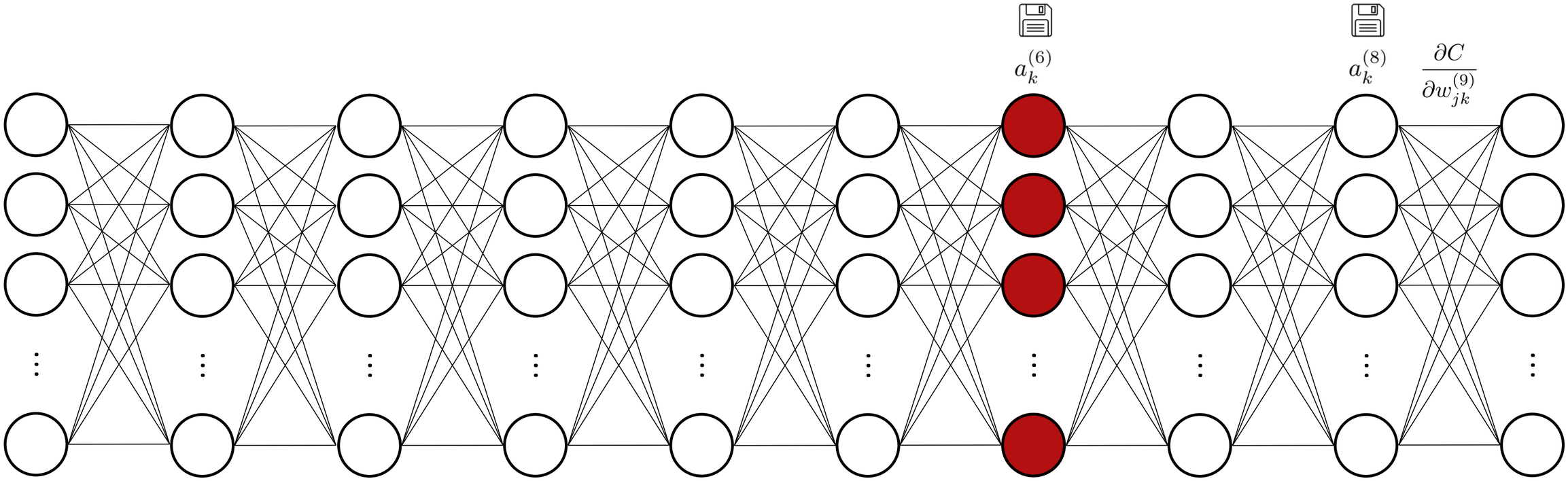


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

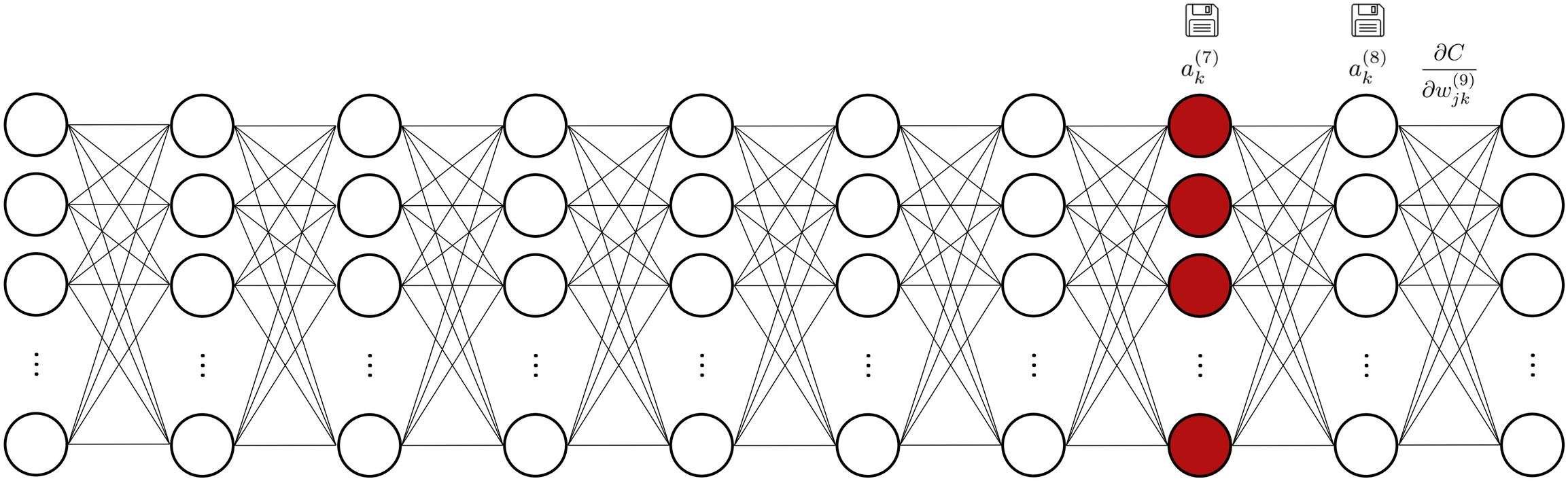


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

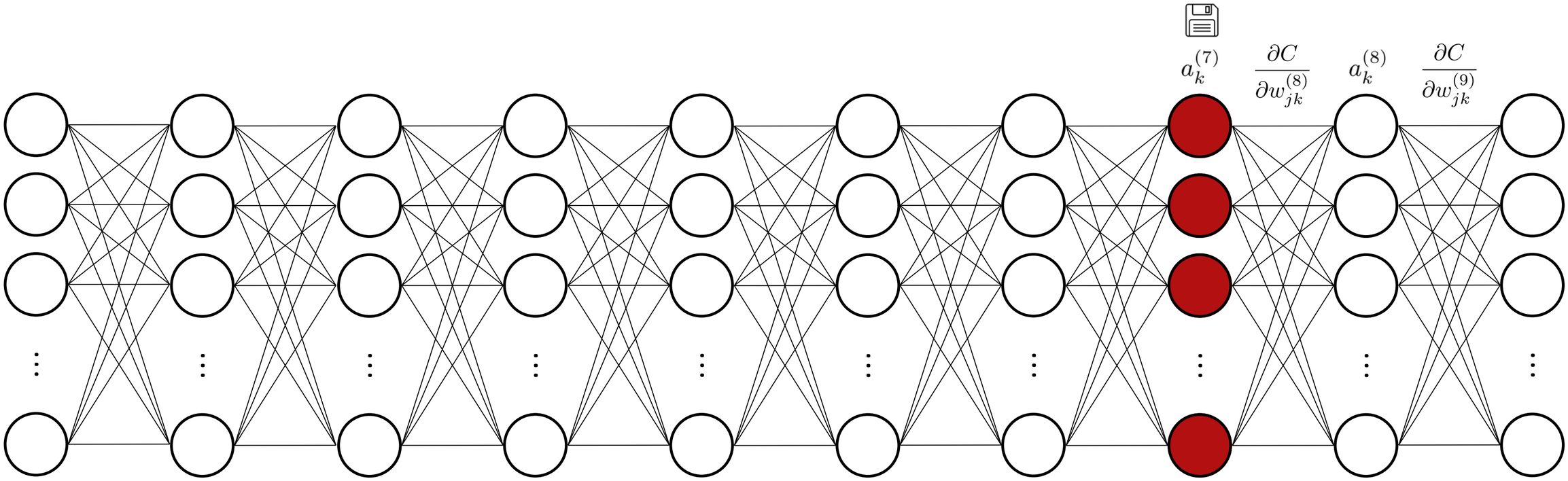


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

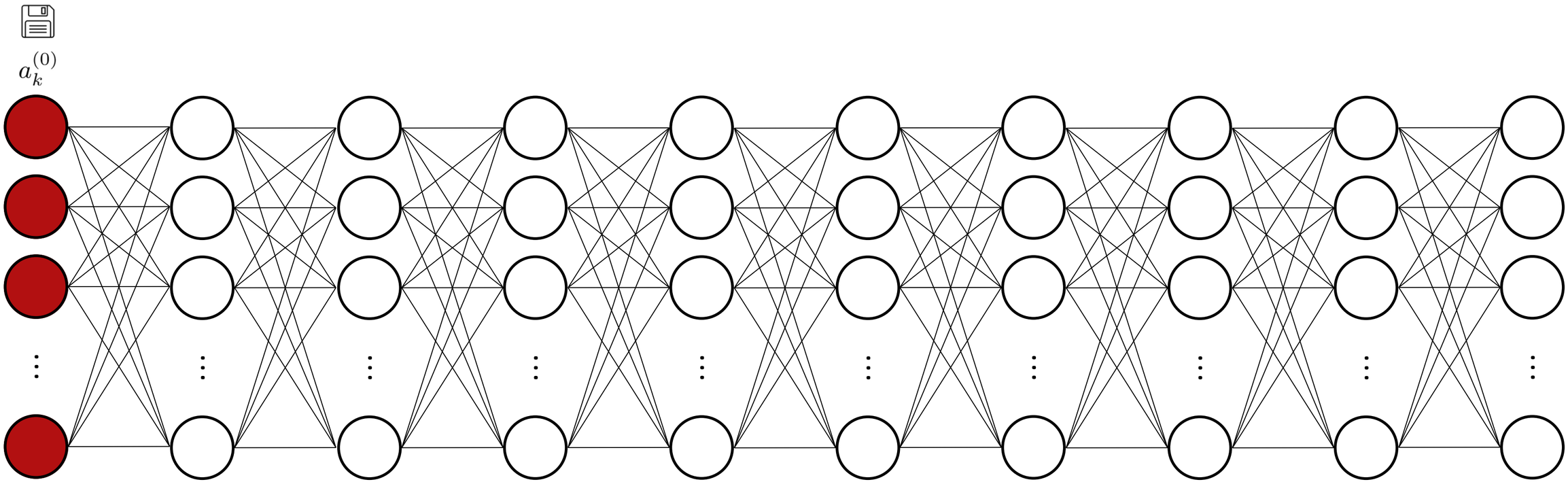


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

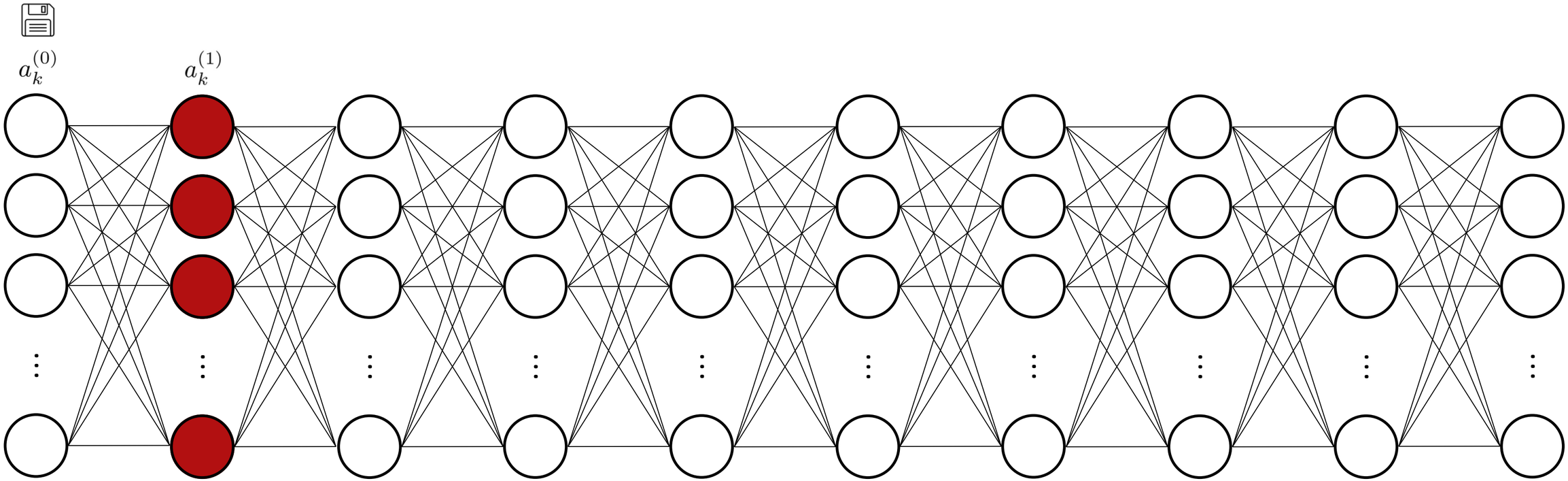


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

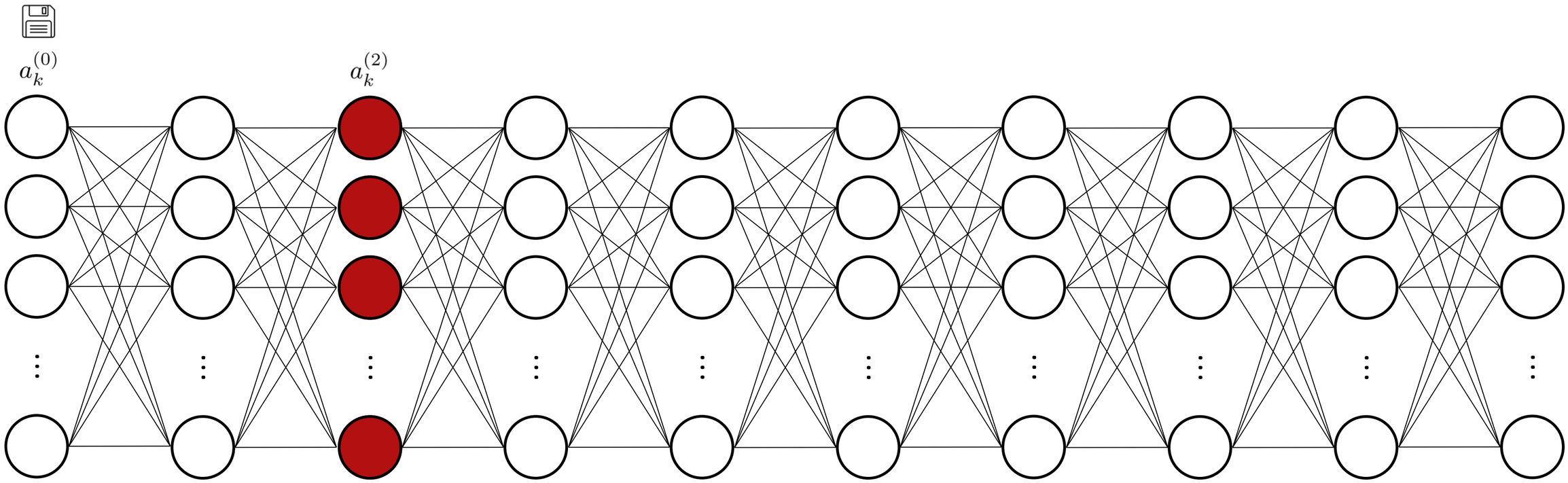


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

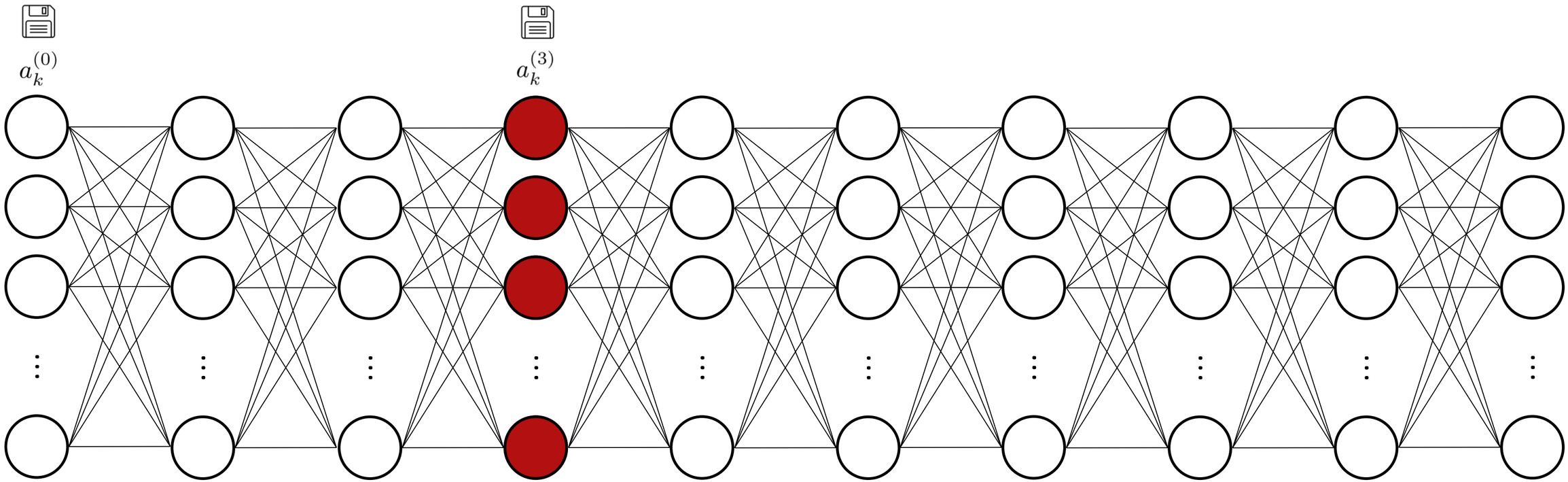


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

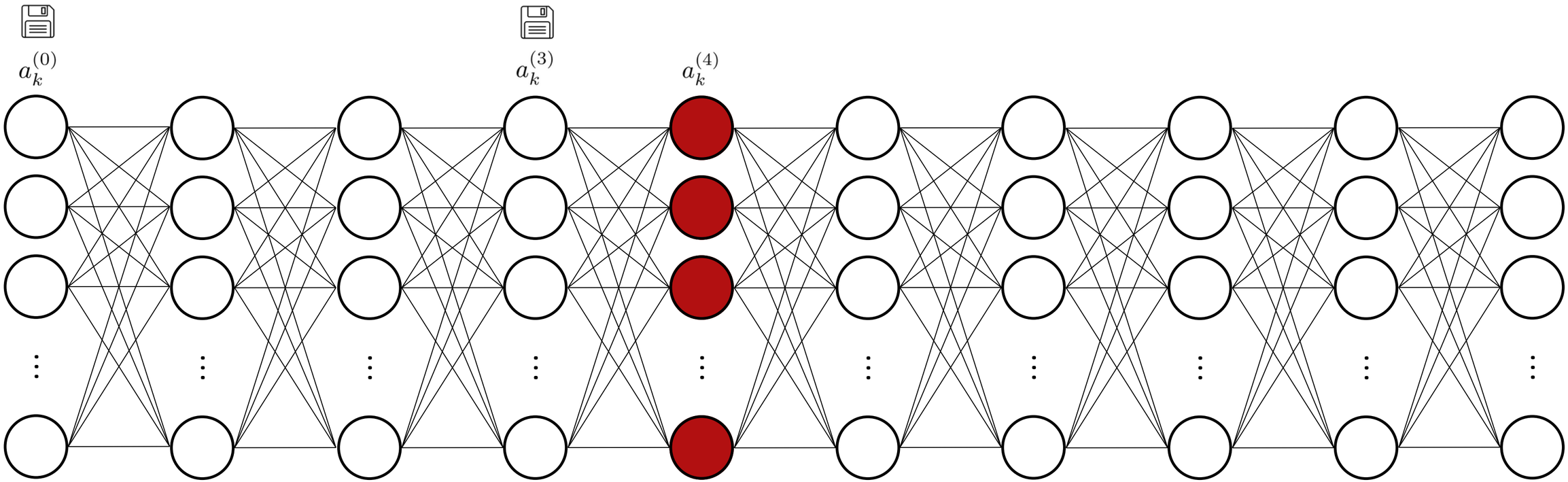


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

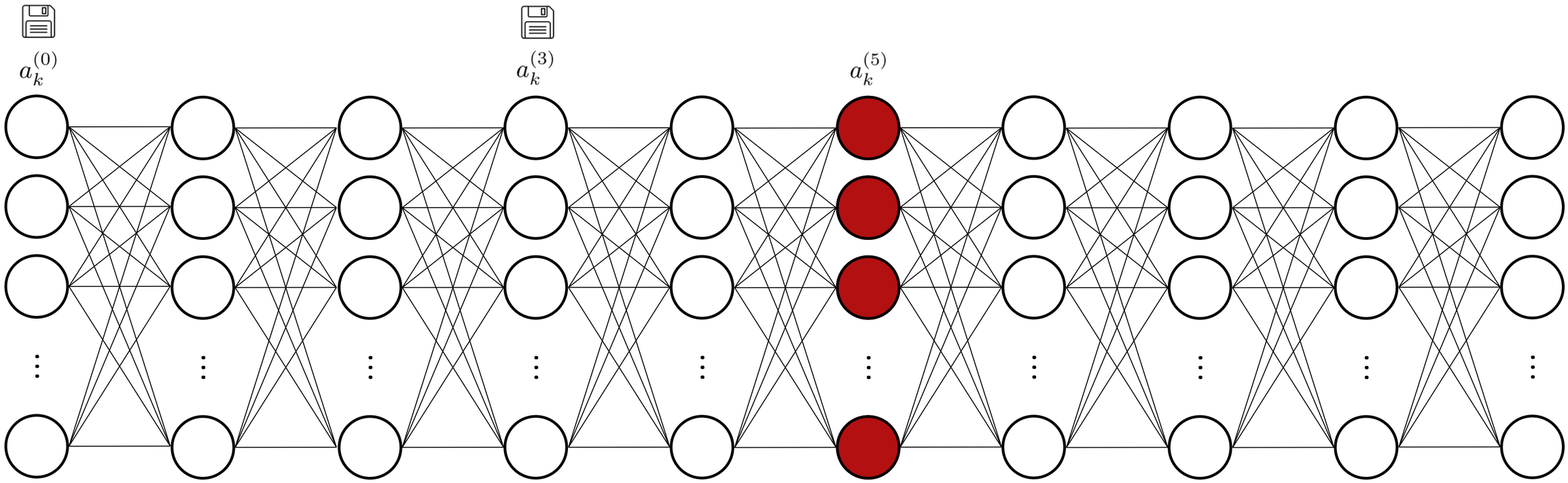


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

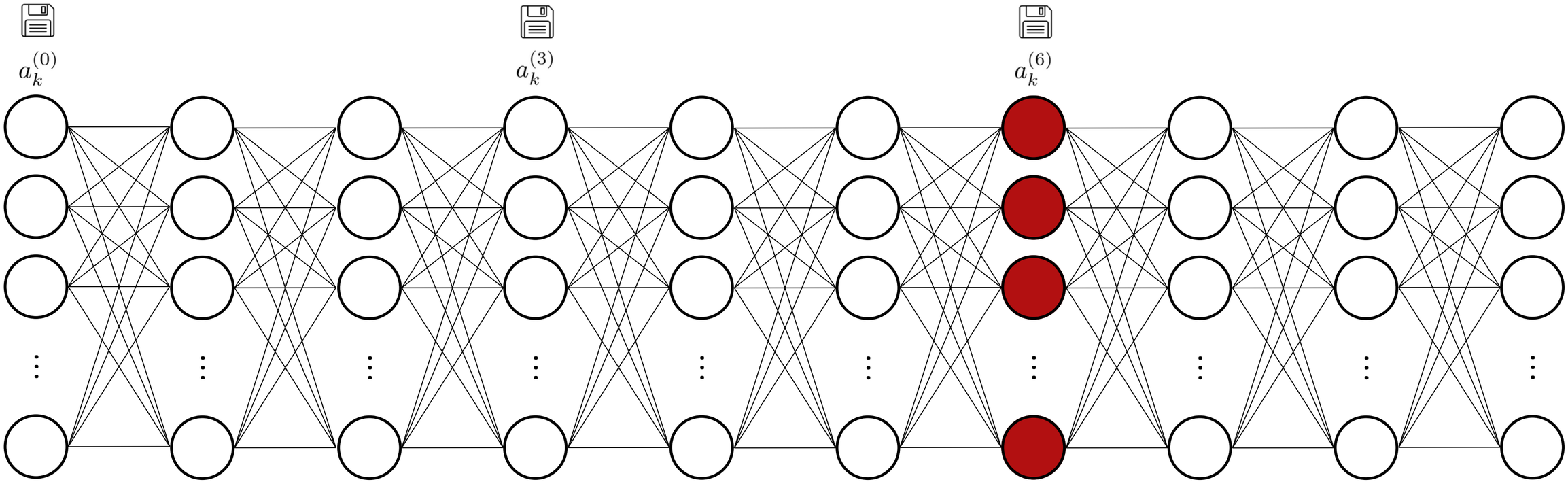


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

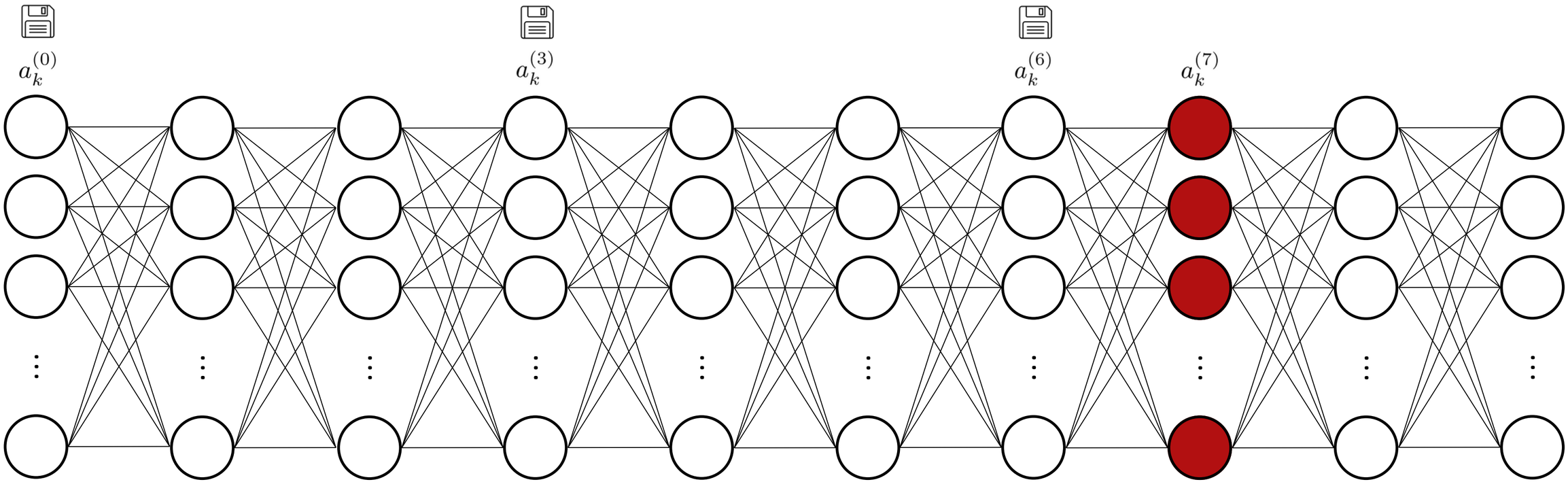


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

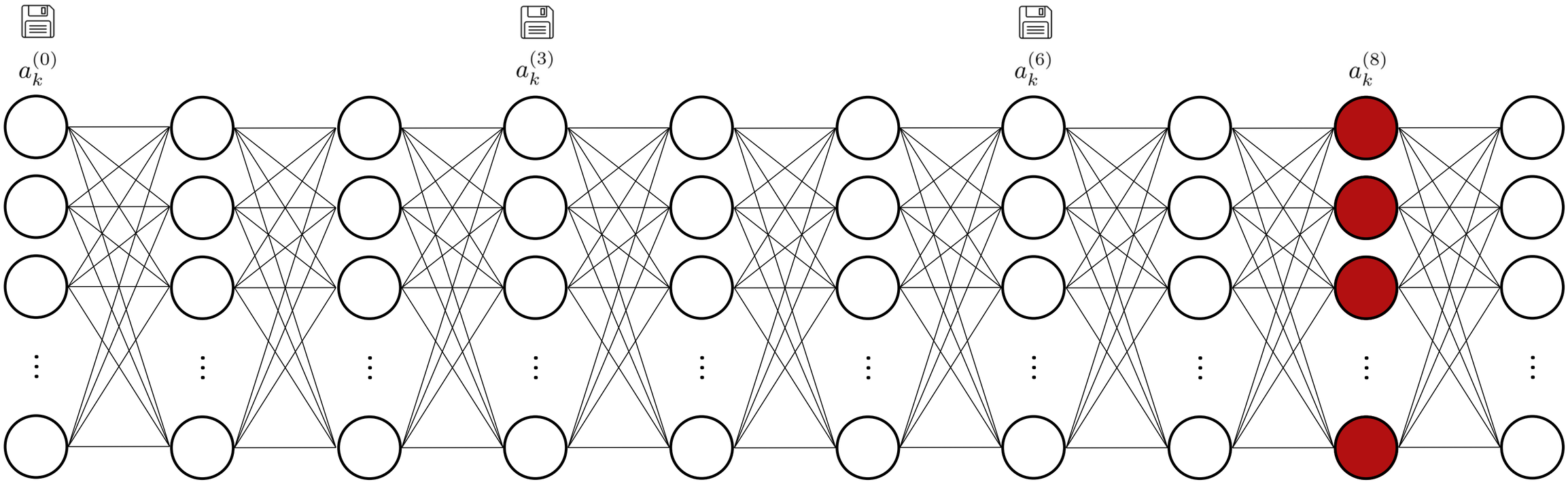


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

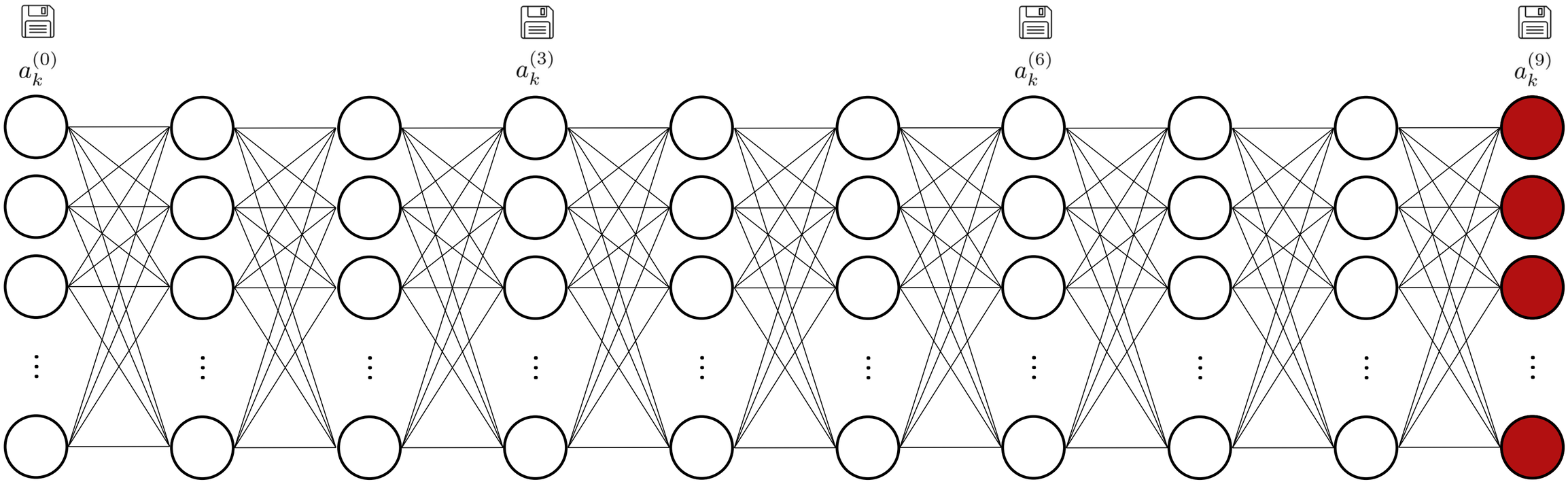


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

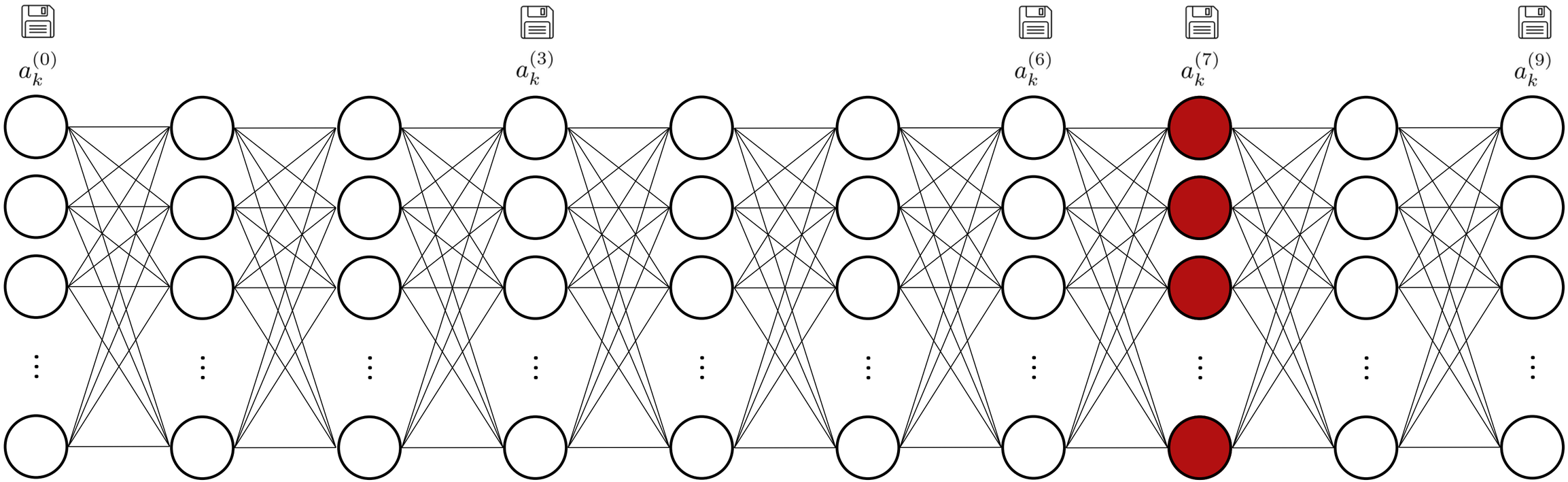


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

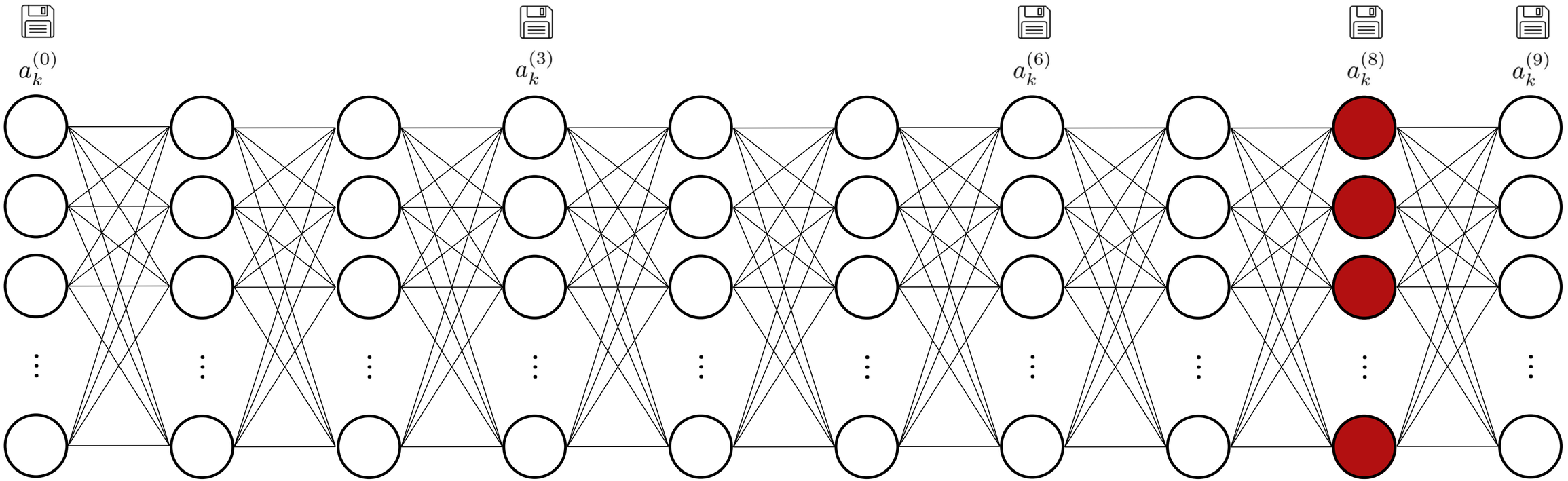


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

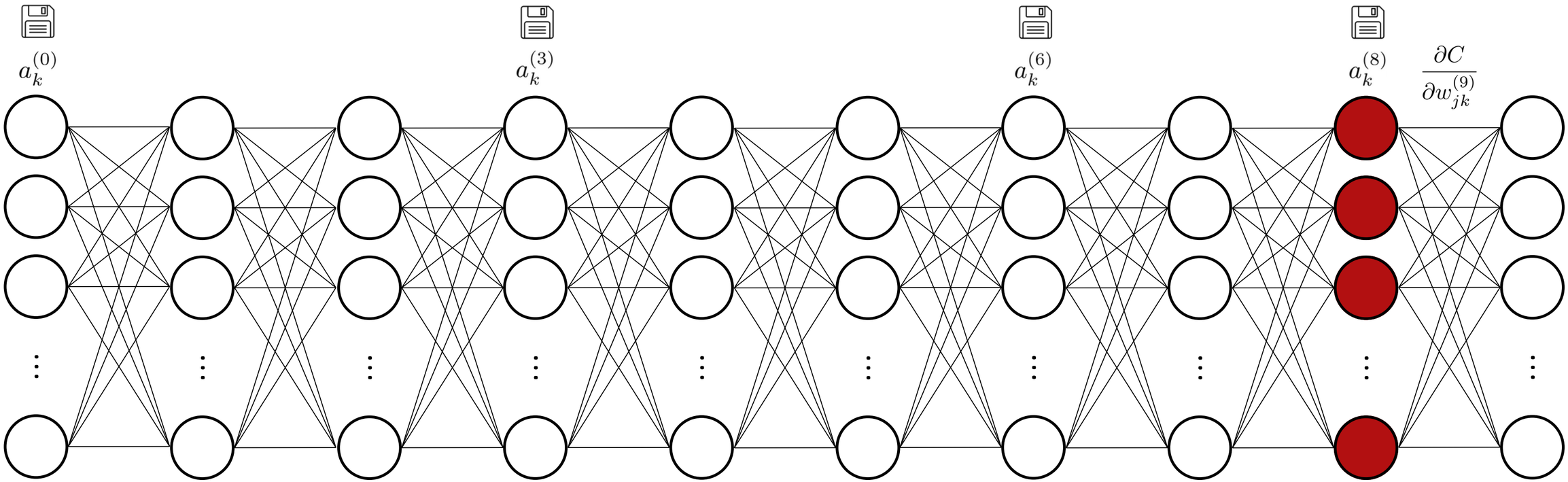


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

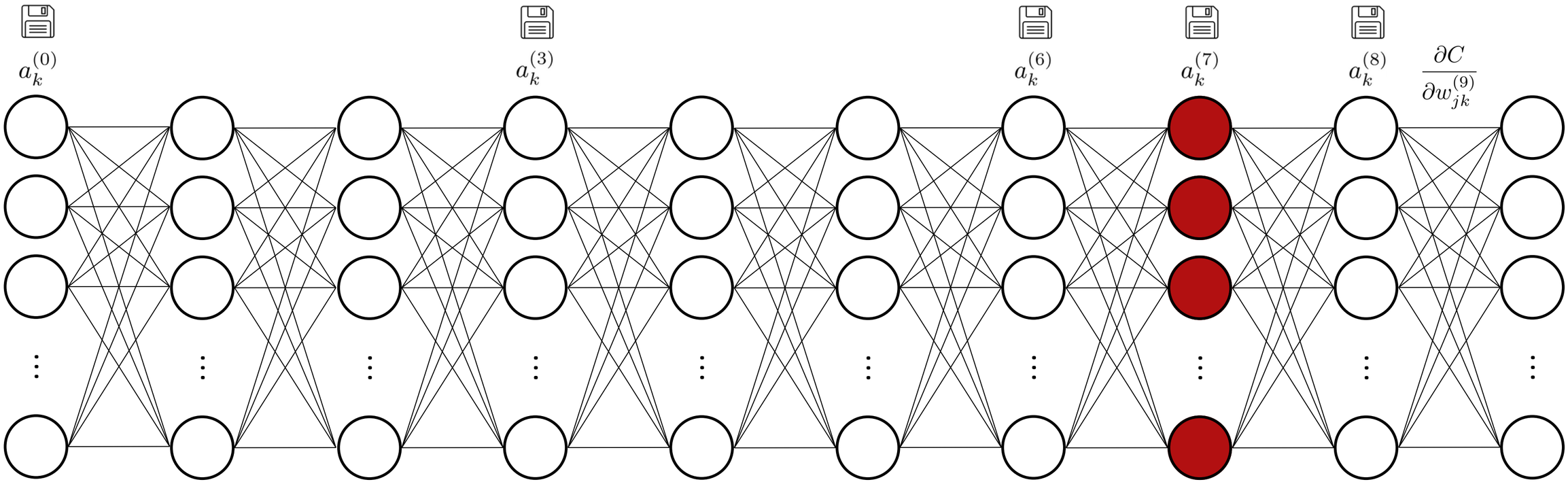


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

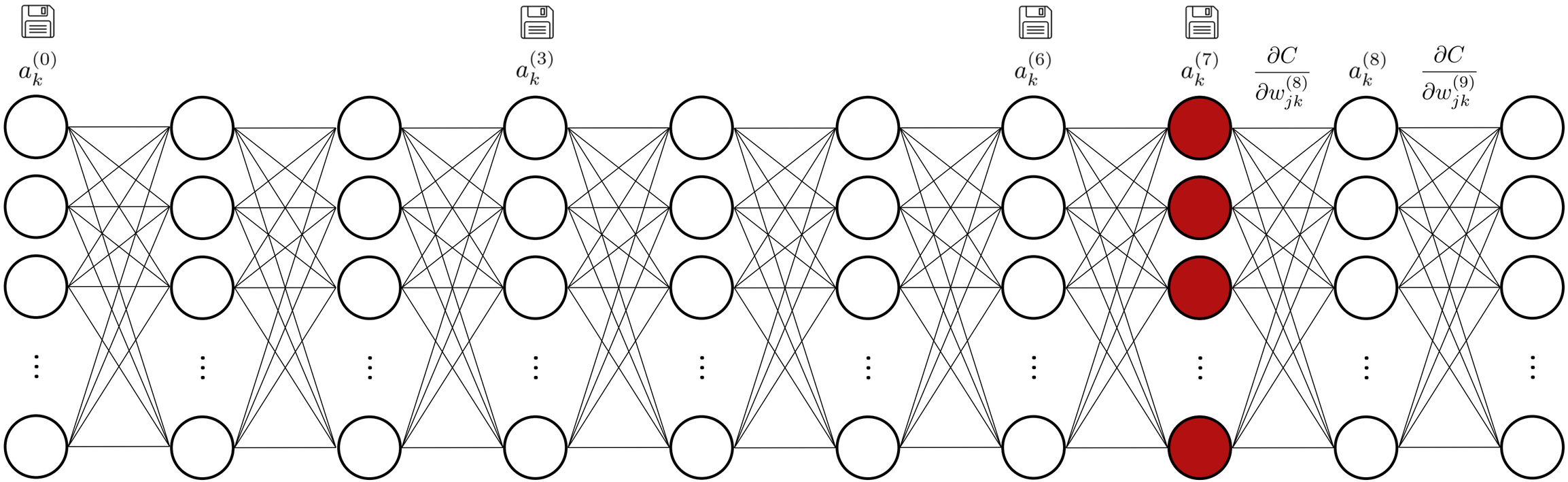


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

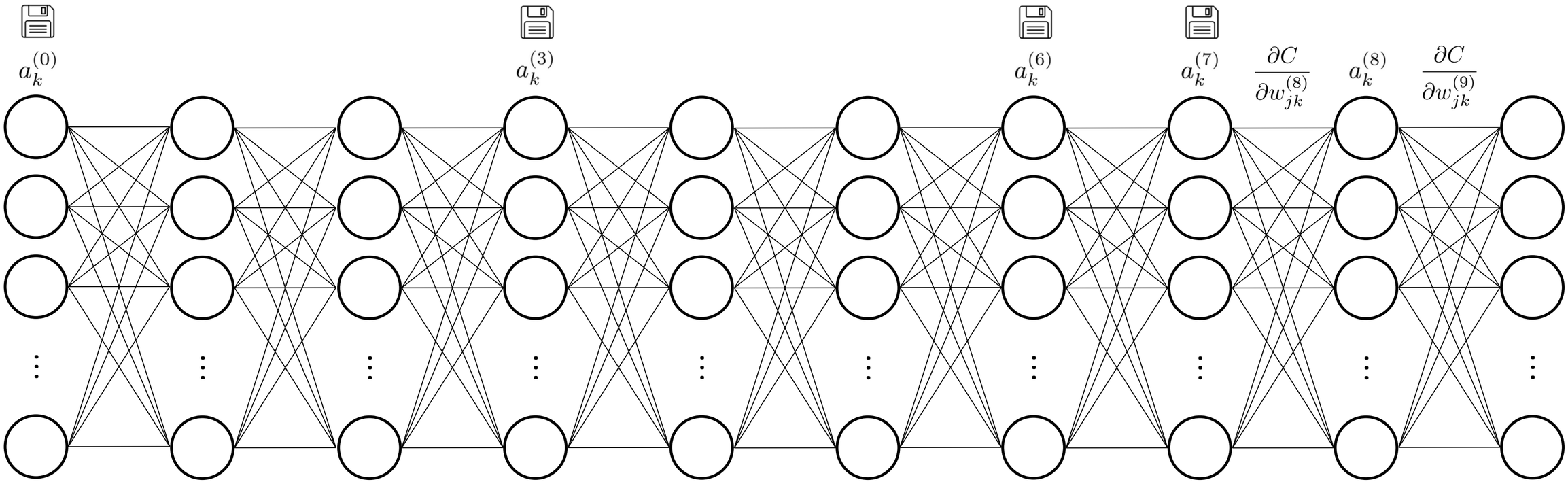


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

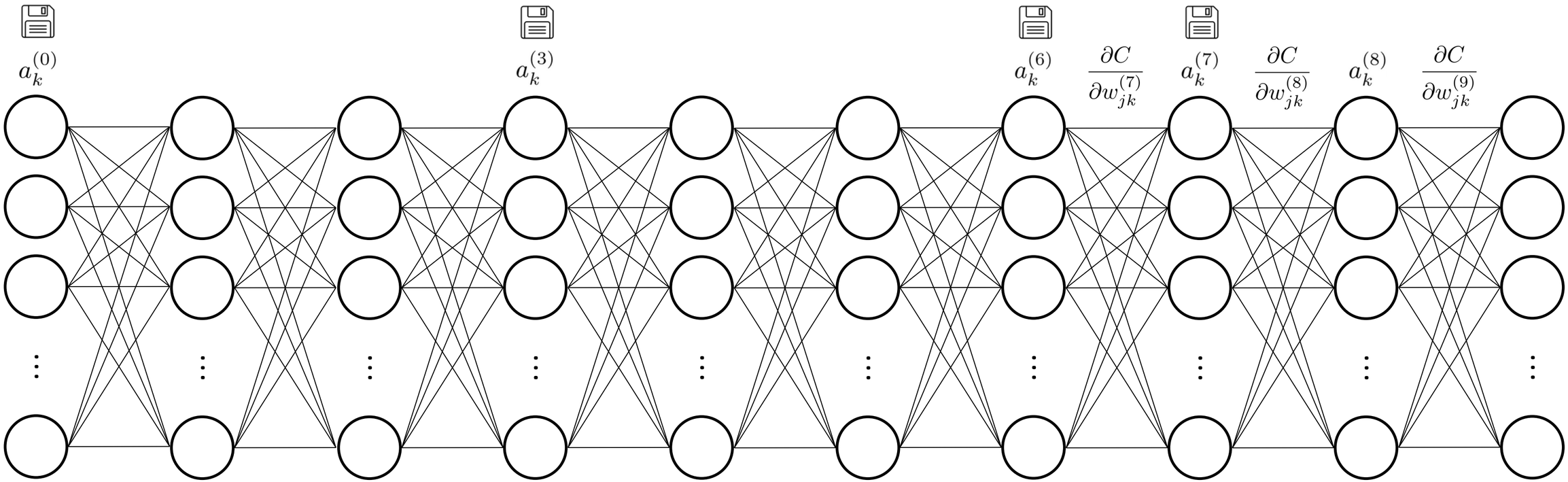


Optimization Stack

Gradient Checkpointing

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$

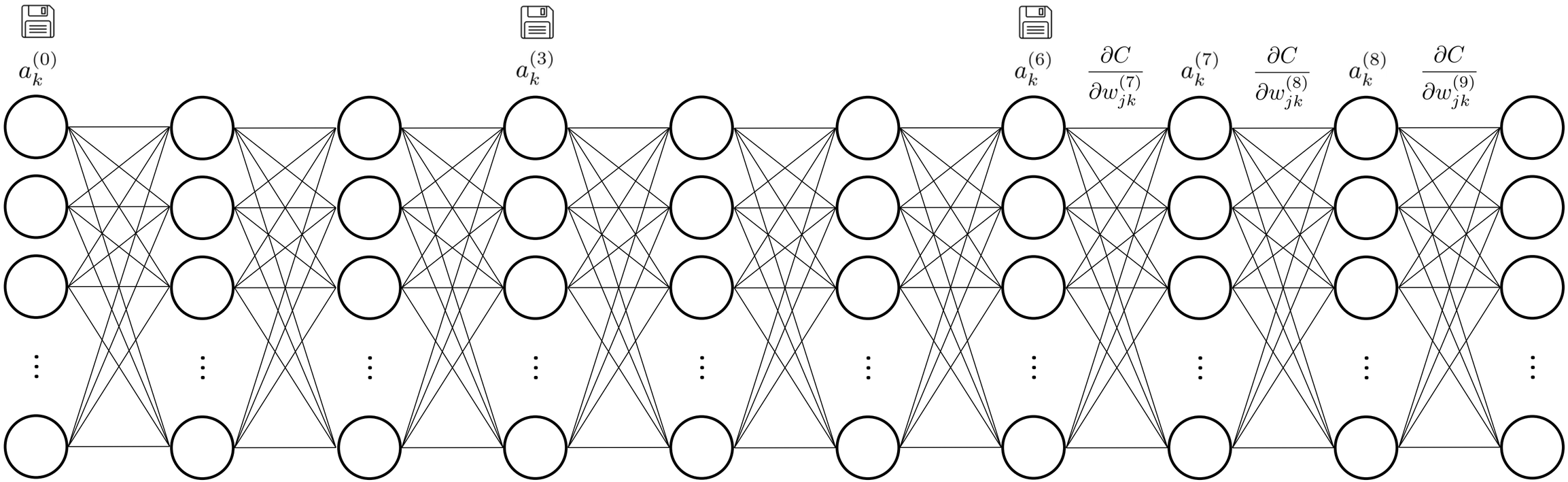


Optimization Stack

Gradient Checkpointing

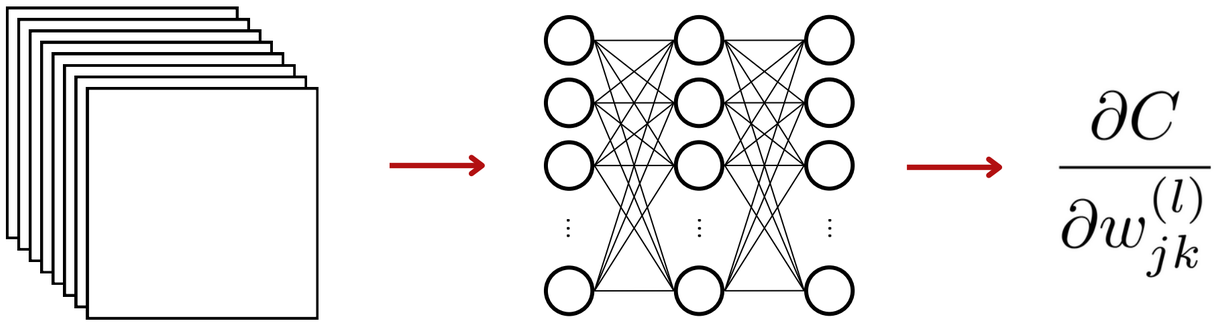
$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)}\, a_k^{(l-1)} \qquad \frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)}\, w_{kj}^{(l+1)}\, a_j^{(l)} \left(1 - a_j^{(l)}\right) \qquad \delta_j^{(L)} = 2\left(y_j - a_j^{(L)}\right) a_j^{(L)} \left(1 - a_j^{(L)}\right)$$



Optimization Stack

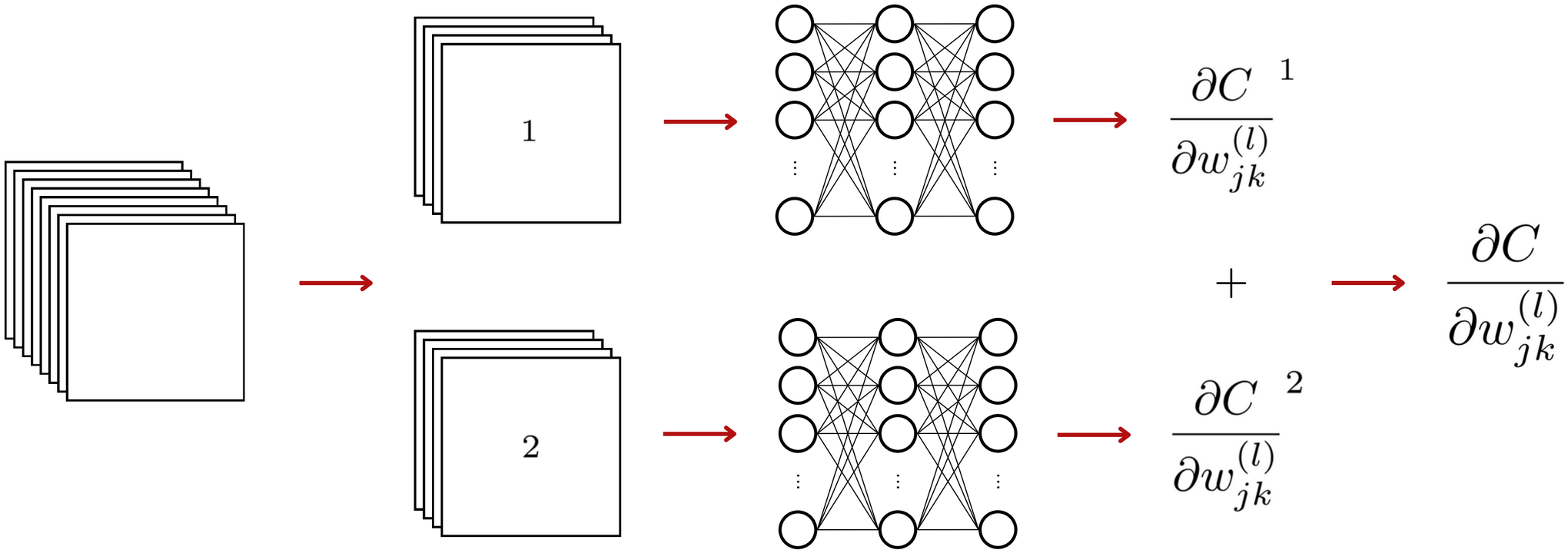
Microbatching



$$\frac{\partial C}{\partial w_{jk}^{(l)}}$$

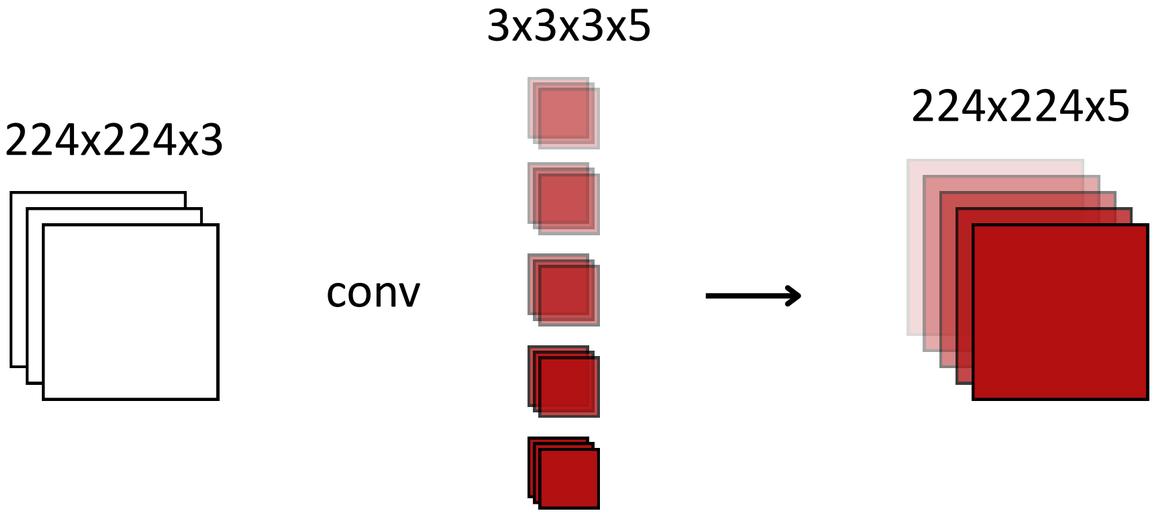
Optimization Stack

Microbatching



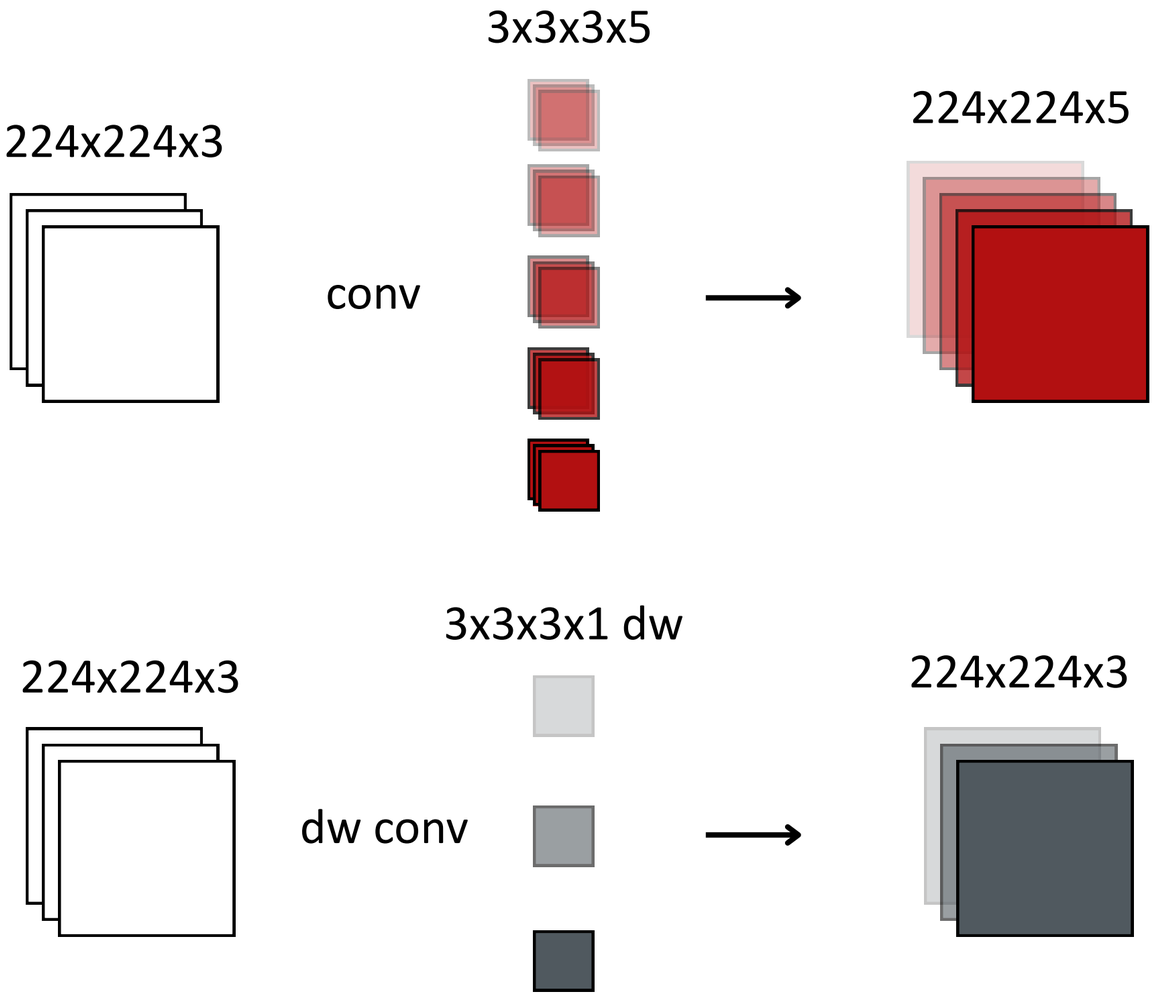
Optimization Stack

MobileNet



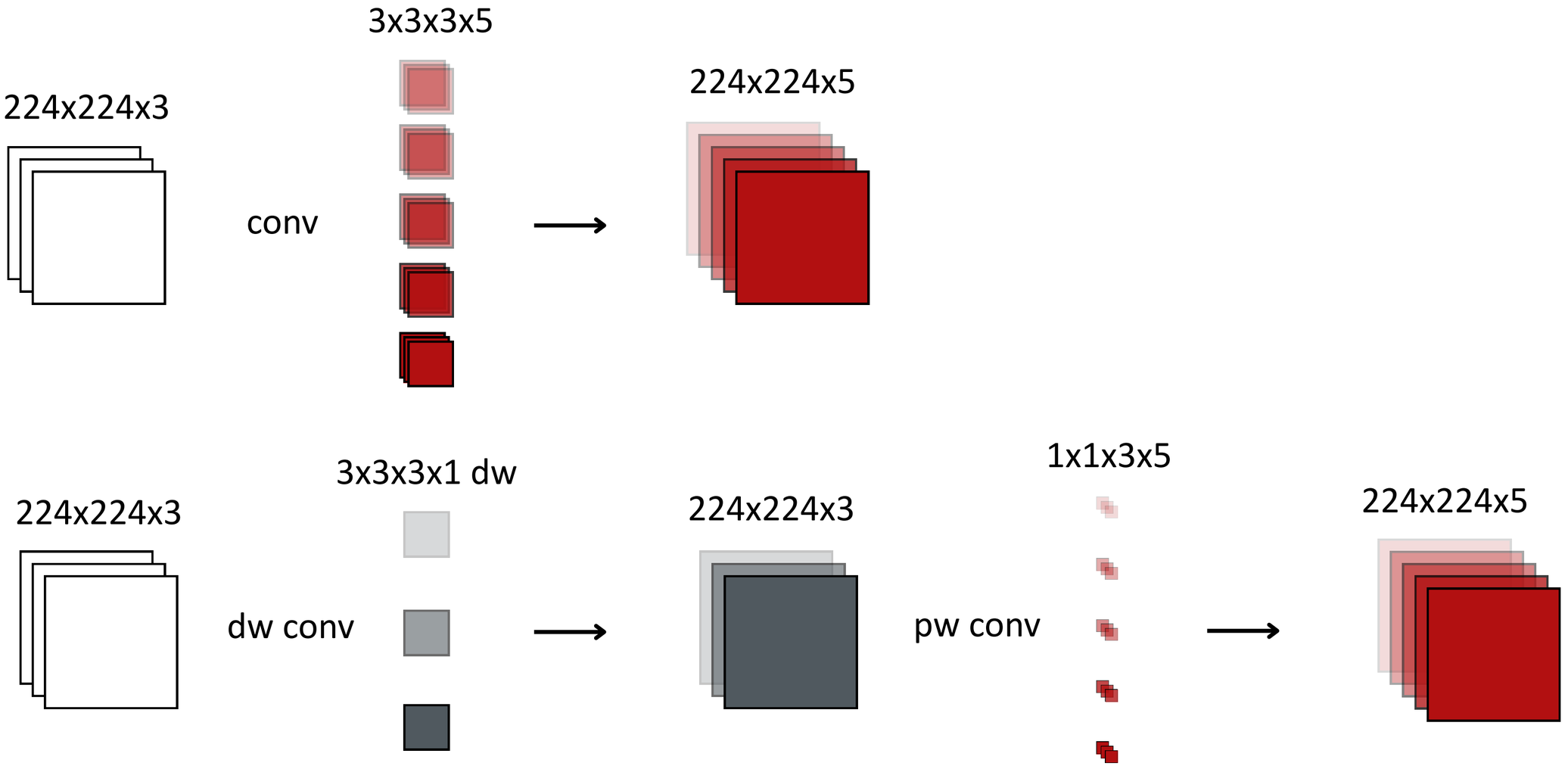
Optimization Stack

MobileNet



Optimization Stack

MobileNet



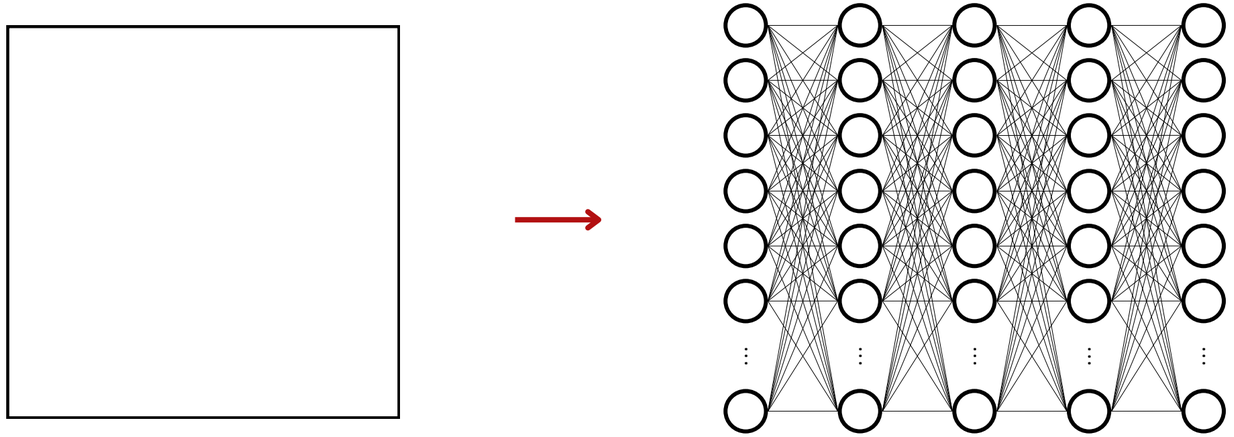
Optimization Stack

MobileNet

Layer	Input Size	Output Size	Filter Shape; Stride
Convolution	$224 \times 224 \times 3$	$112 \times 112 \times 32$	$3, 3, 3, 32; 2$
Depthwise Separable 1	$112 \times 112 \times 32$	$112 \times 112 \times 64$	$3, 3, 32, 1$ dw; 1 $1, 1, 32, 64; 1$
Depthwise Separable 2	$112 \times 112 \times 64$	$56 \times 56 \times 128$	$3, 3, 64, 1$ dw; 2 $1, 1, 64, 128; 1$
Depthwise Separable 3	$56 \times 56 \times 128$	$56 \times 56 \times 128$	$3, 3, 128, 1$ dw; 1 $1, 1, 128, 128; 1$
Depthwise Separable 4	$56 \times 56 \times 128$	$28 \times 28 \times 256$	$3, 3, 128, 1$ dw; 2 $1, 1, 128, 256; 1$
Depthwise Separable 5	$28 \times 28 \times 256$	$28 \times 28 \times 256$	$3, 3, 256, 1$ dw; 1 $1, 1, 256, 256; 1$
Depthwise Separable 6	$28 \times 28 \times 256$	$14 \times 14 \times 512$	$3, 3, 256, 1$ dw; 2 $1, 1, 256, 512; 1$
Depthwise Separable 7–11	$14 \times 14 \times 512$	$14 \times 14 \times 512$	[$3, 3, 512, 1$ dw; 1 $1, 1, 512, 512; 1$] $\times 5$
Depthwise Separable 12	$14 \times 14 \times 512$	$7 \times 7 \times 1024$	$3, 3, 512, 1$ dw; 2 $1, 1, 512, 1024; 1$
Depthwise Separable 13	$7 \times 7 \times 1024$	$7 \times 7 \times 1024$	$3, 3, 1024, 1,$ dw; 1 $1, 1, 1024, 1024; 1$
Average Pool	$7 \times 7 \times 1024$	1024	$7, 7$
Fully Connected	1024	8	$1024, 8$

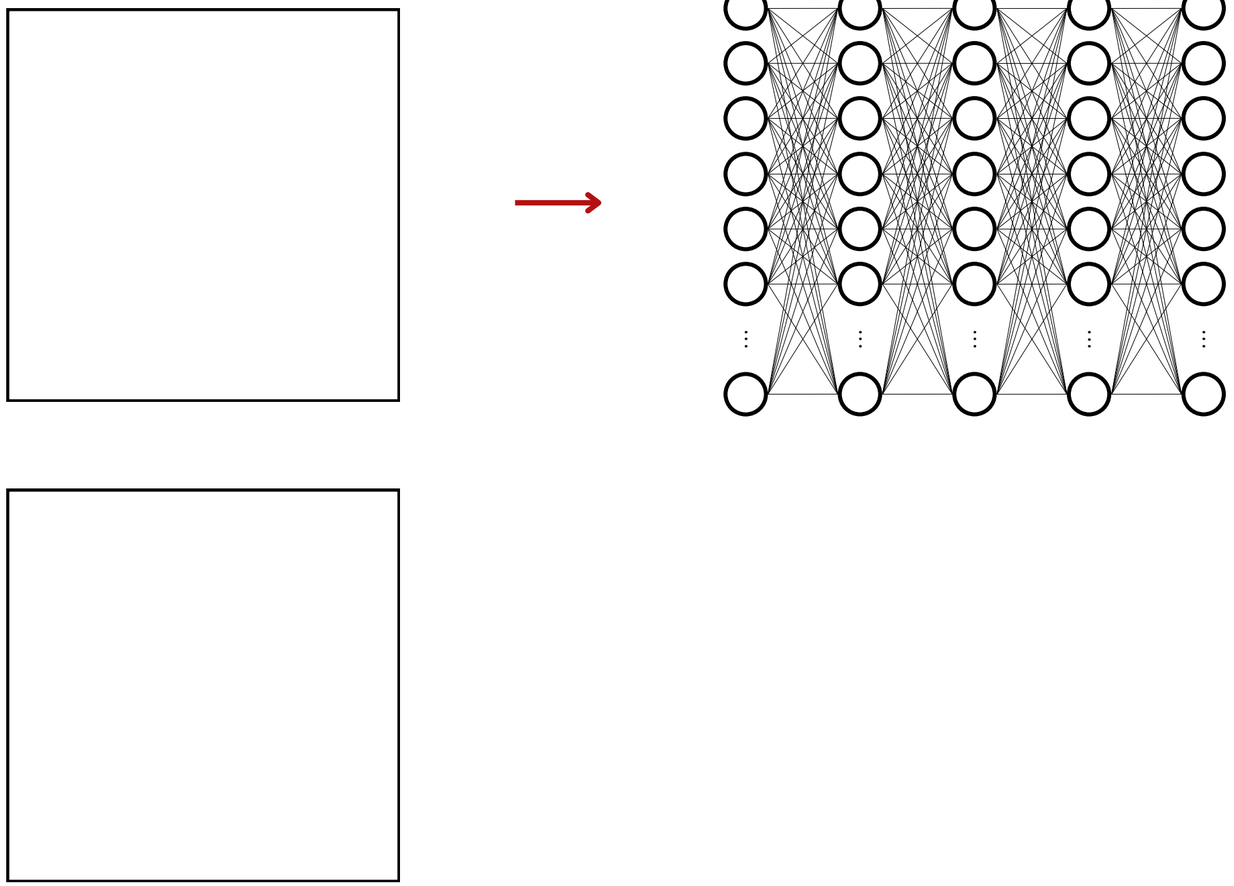
Optimization Stack

Patching



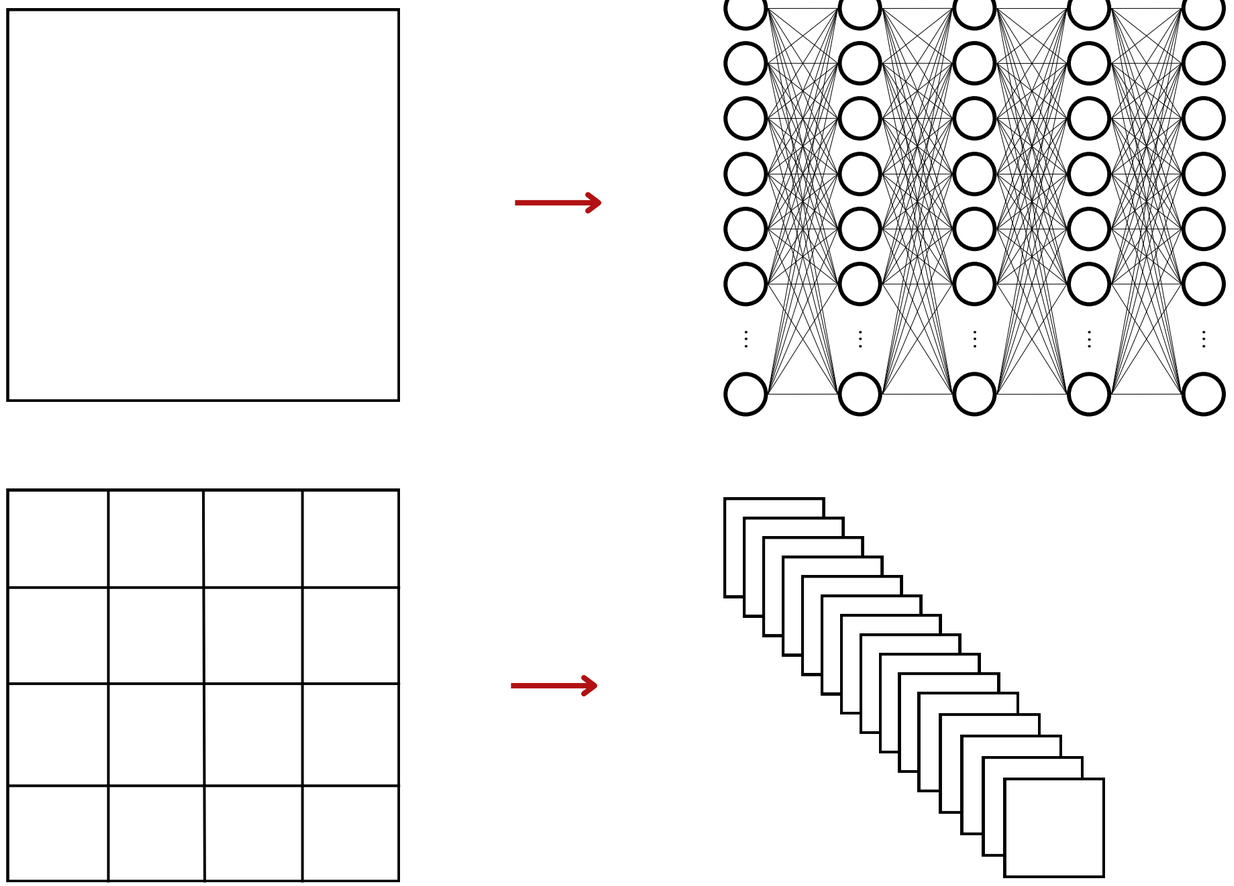
Optimization Stack

Patching



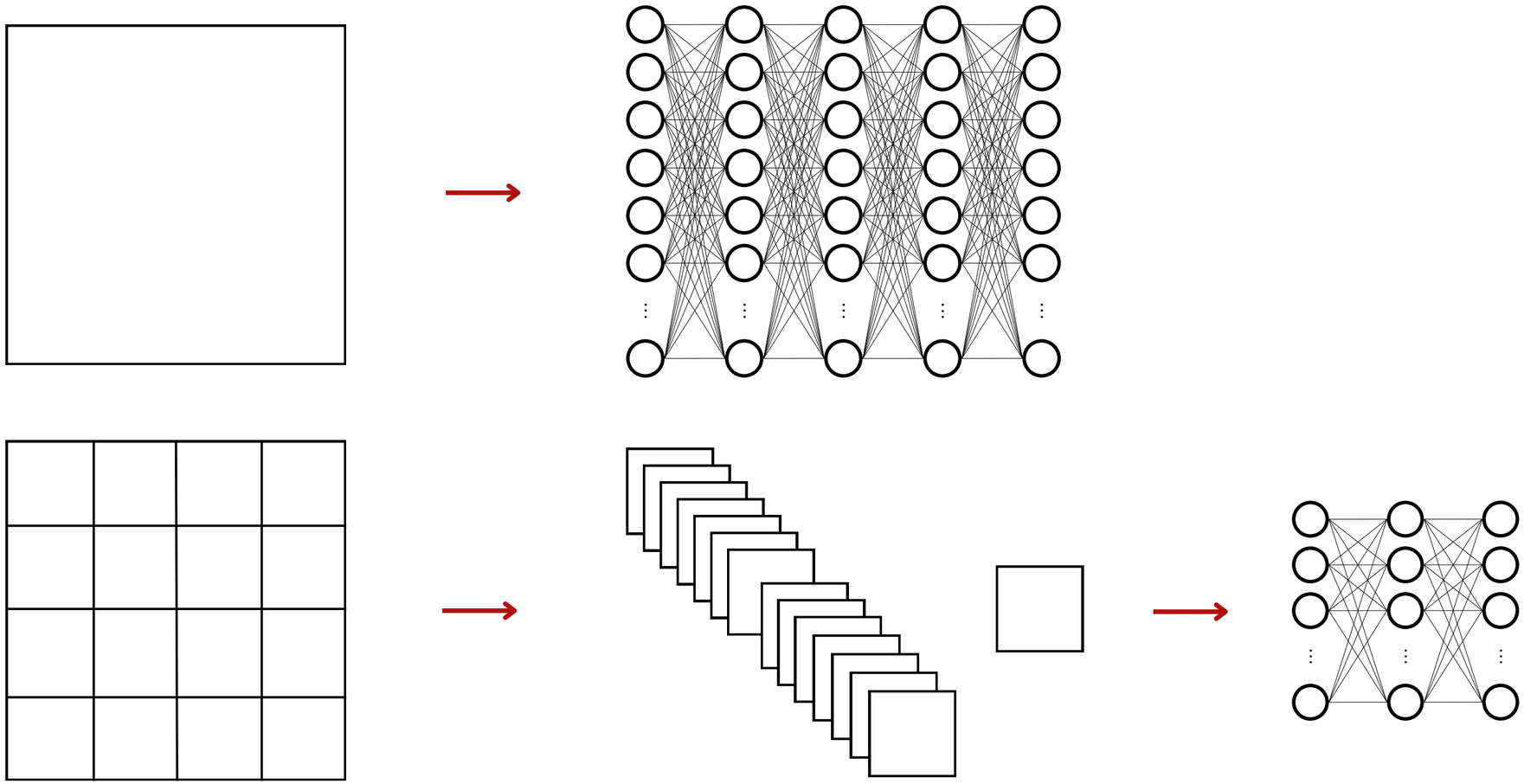
Optimization Stack

Patching



Optimization Stack

Patching

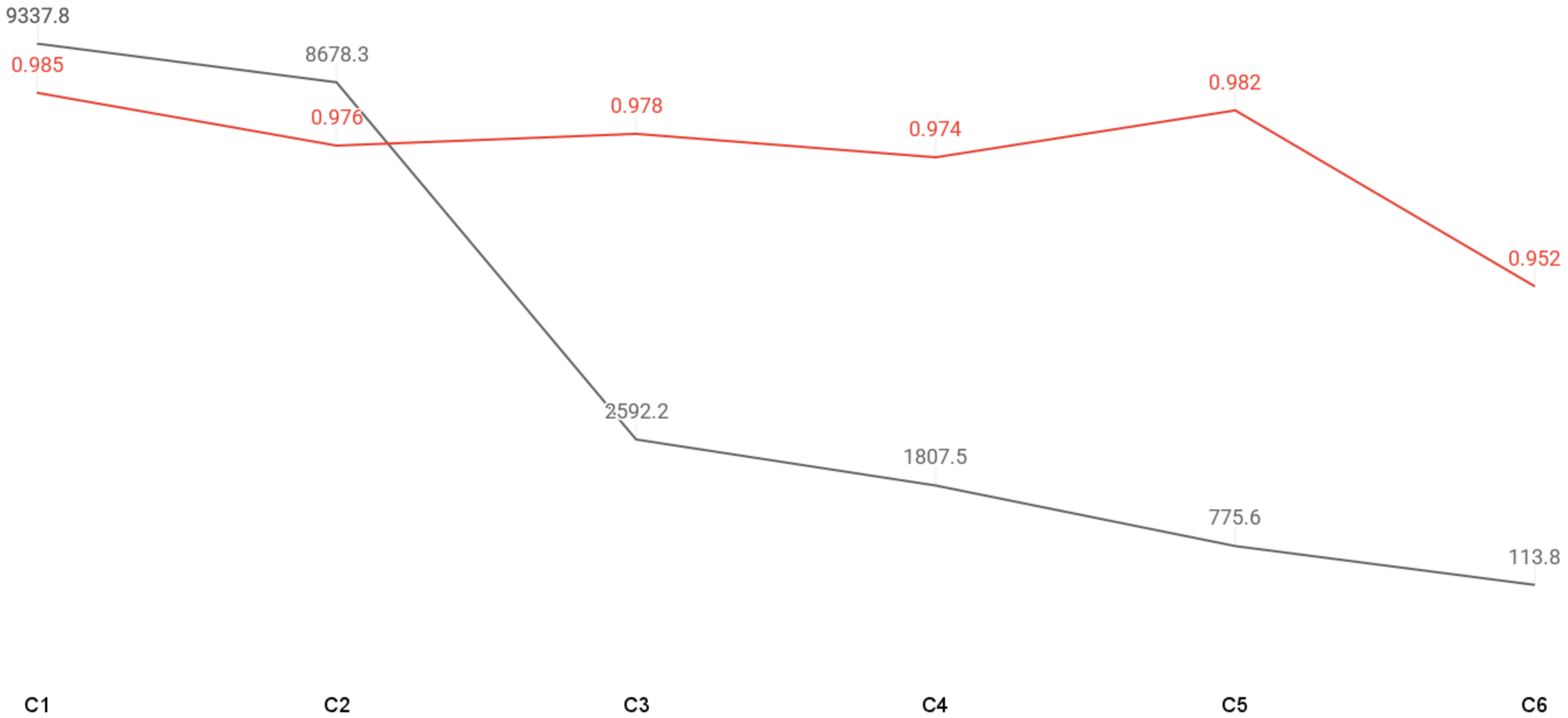


Optimization Stack

Patching

Layer	Input Size	Output Size	Filter Shape; Stride
Convolution	$32 \times 32 \times 3$	$32 \times 32 \times 32$	$3, 3, 3, 32; 1$
Depthwise Separable 1	$32 \times 32 \times 32$	$32 \times 32 \times 64$	$3, 3, 32, 1$ dw; 1 $1, 1, 32, 64; 1$
Depthwise Separable 2	$32 \times 32 \times 64$	$16 \times 16 \times 128$	$3, 3, 64, 1$ dw; 2 $1, 1, 64, 128; 1$
Depthwise Separable 3	$16 \times 16 \times 128$	$16 \times 16 \times 128$	$3, 3, 128, 1$ dw; 1 $1, 1, 128, 128; 1$
Depthwise Separable 4	$16 \times 16 \times 128$	$8 \times 8 \times 256$	$3, 3, 128, 1$ dw; 2 $1, 1, 128, 256; 1$
Depthwise Separable 5	$8 \times 8 \times 256$	$8 \times 8 \times 256$	$3, 3, 256, 1$ dw; 1 $1, 1, 256, 256; 1$
Depthwise Separable 6	$8 \times 8 \times 256$	$4 \times 4 \times 512$	$3, 3, 256, 1$ dw; 2 $1, 1, 256, 512; 1$
Depthwise Separable 7	$4 \times 4 \times 512$	$4 \times 4 \times 1024$	$3, 3, 512, 1$ dw; 1 $1, 1, 512, 1024; 1$
Depthwise Separable 8	$4 \times 4 \times 1024$	$4 \times 4 \times 1024$	$3, 3, 1024, 1$ dw; 1 $1, 1, 1024, 1024; 1$
Average Pool	$4 \times 4 \times 1024$	1024	$4, 4$
Fully Connected	1024	8	$1024, 8$

Results


9337.8
8678.3
2592.2
1807.5
775.6
113.8
0.985
0.976
0.978
0.974
0.982
0.952
C1
C2
C3
C4
C5
C6

Thank You