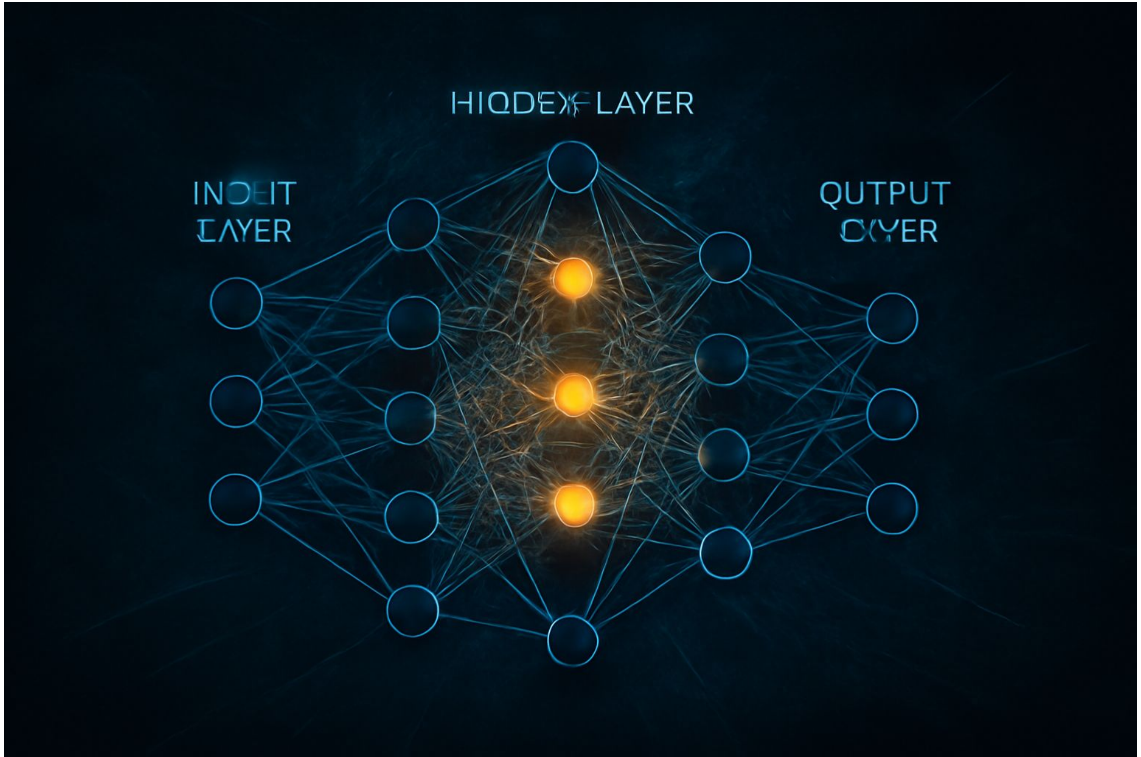
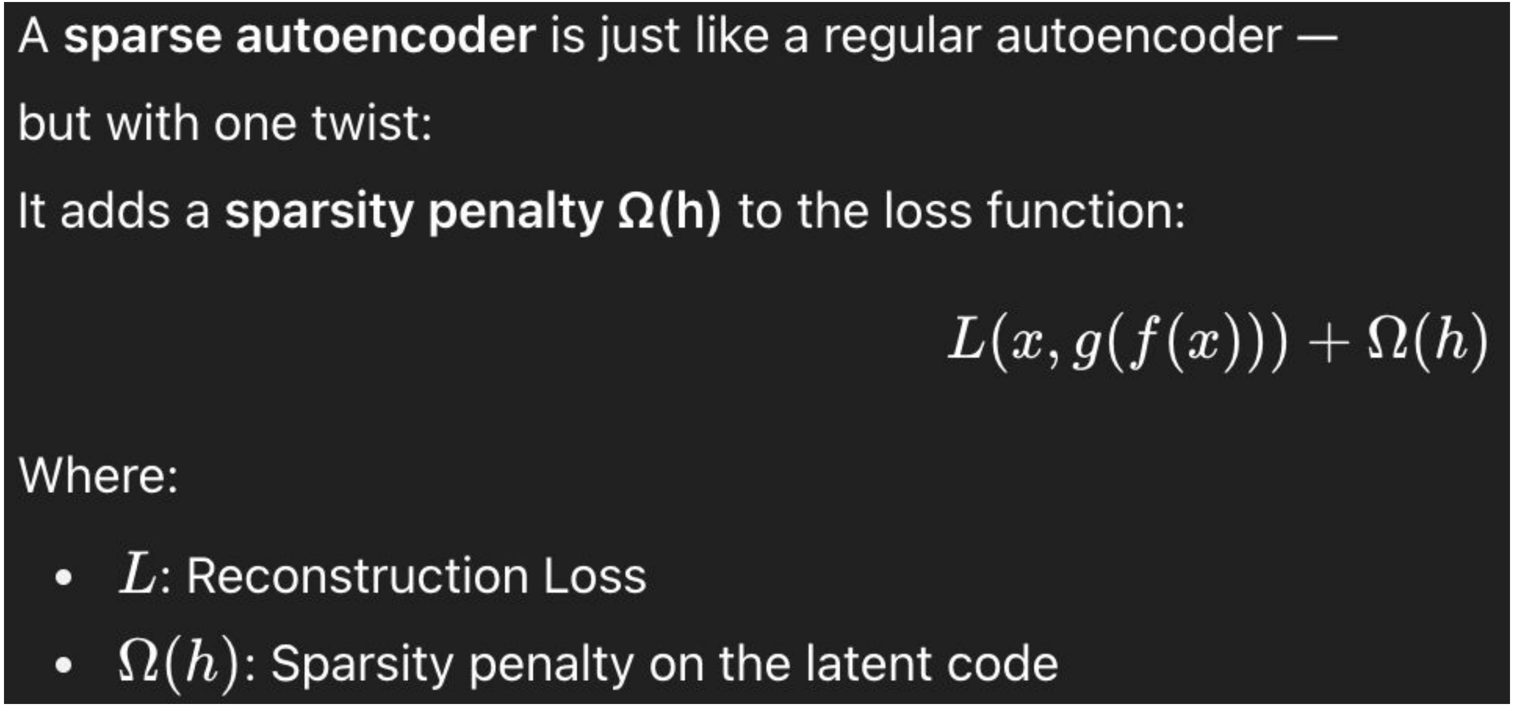
Sparse Autoencoders

Learning Useful Features Through Sparsity



What is a Sparse Autoencoder?



A **sparse autoencoder** is just like a regular autoencoder — but with one twist:

It adds a **sparsity penalty Ω(h)** to the loss function:

$$L(x, g(f(x))) + \Omega(h)$$

Where:

- L: Reconstruction Loss
- $\Omega(h)$: Sparsity penalty on the latent code

Analogy — Summarizing a Book

Imagine **summarizing a 500-page book with just 3 bullet points.**

You’re forced to pick the most important ideas.

That’s what sparsity does — **only the most important features “survive.”**

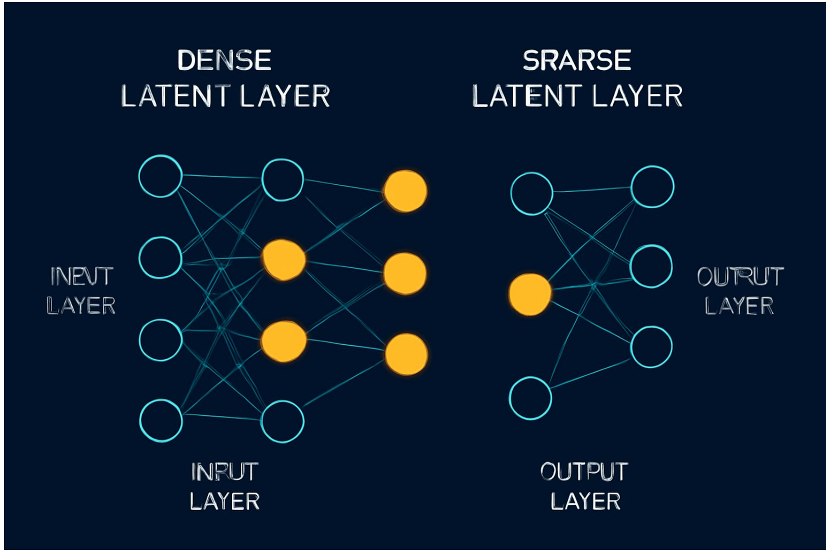
Why Do We Need Sparsity?

Without sparsity:

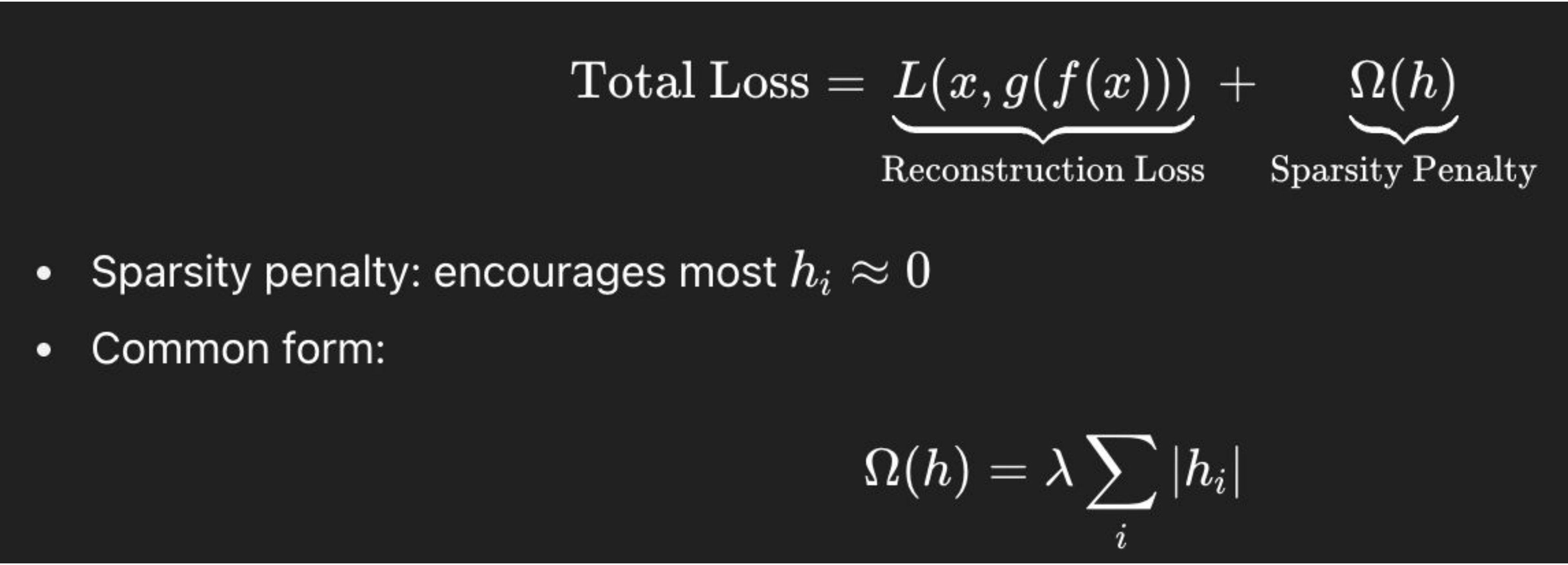
- Autoencoders may just **copy inputs** = identity function

With sparsity:

- Must learn **unique, statistical features of the data**
- Learns useful features even without labels



The Loss Function Revisited



$$\text{Total Loss} = \underbrace{L(x, g(f(x)))}_{\text{Reconstruction Loss}} + \underbrace{\Omega(h)}_{\text{Sparsity Penalty}}$$

- Sparsity penalty: encourages most $h_i \approx 0$
- Common form:

$$\Omega(h) = \lambda \sum_i |h_i|$$

Not Just Another Regularizer

Weight decay = prior over **parameters**

Sparsity = preference over **functions / representations**

Not Bayesian in the usual sense → **depends on data**

Why Use Sparse Autoencoders?

✅ Learn useful features from unlabeled data

✅ Avoids identity-function problem

✅ Bridges feature learning and generative modeling

✅ Enables interpretability and compression

Quiz Time!

Q1. Why do we add a sparsity penalty in a Sparse Autoencoder?

A. To reduce training time

B. To encourage the autoencoder to memorize the input

C. To ensure only a few neurons in the code layer are active

D. To make the model more complex

Quiz Time!

Q2. Which real-world analogy best describes the role of sparsity in a sparse autoencoder?

A. Memorizing a book word for word

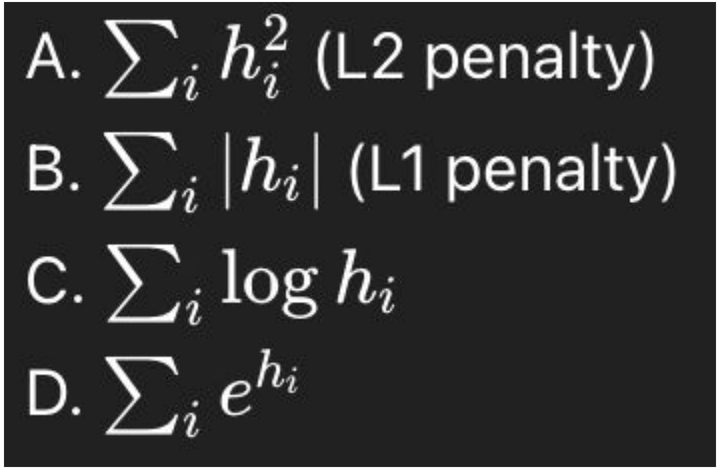
B. Summarizing a book using only 3 bullet points

C. Reading every detail of a newspaper

D. Copying your friend’s homework exactly

Quiz Time!

Q3. Which of the following is a typical form of the sparsity penalty used in sparse autoencoders?



A. $\sum_i h_i^2$ (L2 penalty)

B. $\sum_i |h_i|$ (L1 penalty)

C. $\sum_i \log h_i$

D. $\sum_i e^{h_i}$

Quiz Time!

Q4. What is a potential benefit of using sparse autoencoders for downstream tasks like classification?

A. It increases the size of the dataset

B. It forces overfitting

C. It learns disentangled and interpretable features

D. It ignores unique statistical features of the data