COPPE
UFRJ

Deep Learning

Restricted Boltzmann Machines and Deep Belief Networks

Guilherme Mota

Electrical Engineering Program
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
Universidade Federal do Rio de Janeiro

05/11/2025

Overview

COPPE
UFRJ

1. Key Concepts

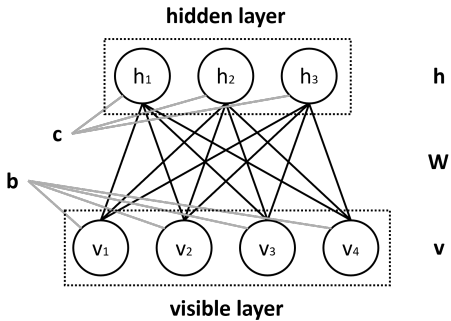
2. Code Implementations

3. Test Results

COPPE
UFRJ

Key Concepts

The RBM



For a binary RBM:

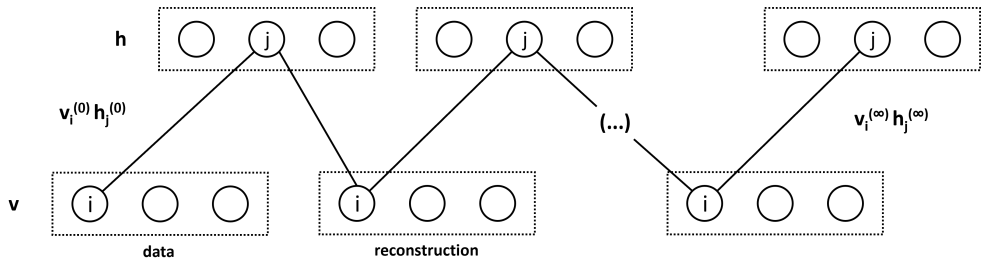
$$\mathbb{P}(h_j \mid \mathbf{v}) = \sigma(c_j + \mathbf{v}^\intercal \mathbf{W}_{:j}) \quad (1)$$

$$\mathbb{P}(v_i \mid \mathbf{h}) = \sigma(b_i + \mathbf{W}_{i:}\mathbf{h}) \quad (2)$$

The RBM Sampling

COPPE
UFRJ

$$\mathbb{P}(v_i \mid \mathbf{h}) = \sigma(b_i + \mathbf{W}_{i:}\mathbf{h})$$



$$\mathbb{P}(h_j \mid \mathbf{v}) = \sigma(c_j + \mathbf{v}^\intercal \mathbf{W}_{:j})$$

The RBM

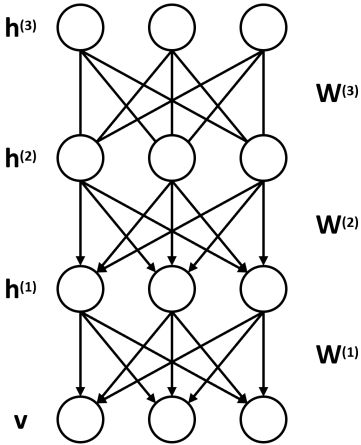
The gradients of loss function with respect to each parameter becomes:

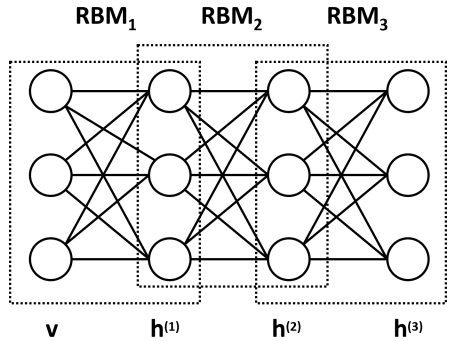
$$\nabla_{\mathbf{W}} \ell(\theta) = \sum_{i=1}^{n} \mathbf{v}_i \cdot \hat{\mathbf{h}}^{\mathsf{T}} - \sum_{i=1}^{n} \tilde{\mathbf{v}}_i \tilde{\mathbf{h}}_i^{\mathsf{T}} \tag{3}$$

$$\nabla_{\mathbf{b}} \ell(\theta) = \sum_{i=1}^{n} \mathbf{v}_i - \sum_{i=1}^{n} \tilde{\mathbf{v}}_i \tag{4}$$

$$\nabla_{\mathbf{c}} \ell(\theta) = \sum_{i=1}^{n} \hat{\mathbf{h}}^{\mathsf{T}} - \sum_{i=1}^{n} \tilde{\mathbf{h}}_i^{\mathsf{T}} \tag{5}$$

The DBN





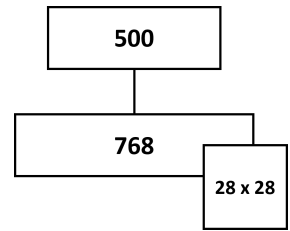
Code Implementations

Test Results

(1 RBM) Binary vs. Probability

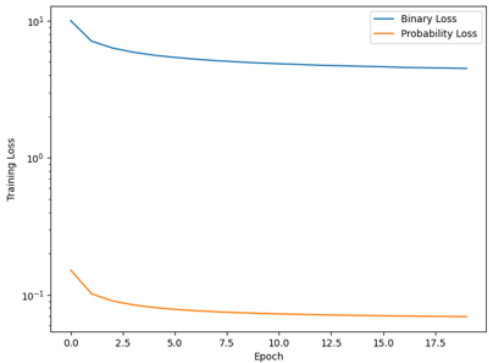
COPPE
UFRJ

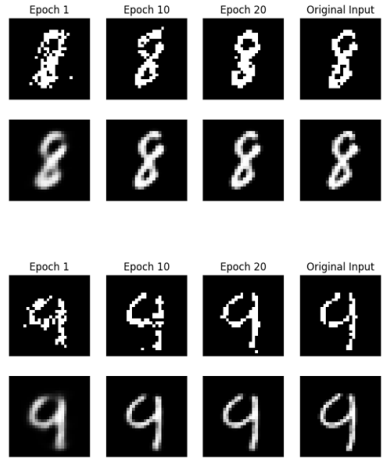
- **Objective**: Evaluate RBM with binary layers vs. probabilistic layers:
  - Binary layers use Bernoulli distribution
  - Probability layers consider the true values sampled from the conditional distributions
- Binary cross entropy as loss function
- Datasets of MNIST and Fashion MNIST:
  - 60.000 (Train) and 10.000 (Test) instances



(1 RBM) MNIST

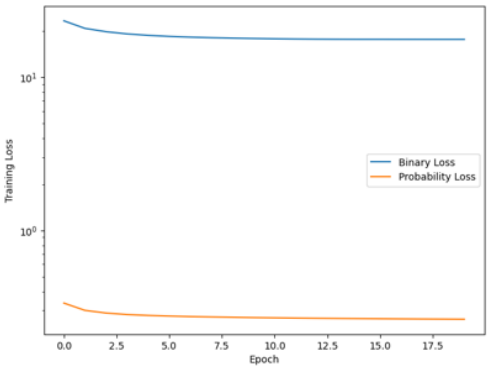
COPPE
UFRJ

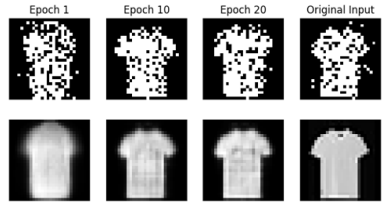


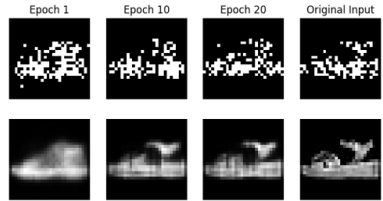


(1 RBM) Fashion MNIST

COPPE
UFRJ







(1 RBM) Results

COPPE
UFRJ

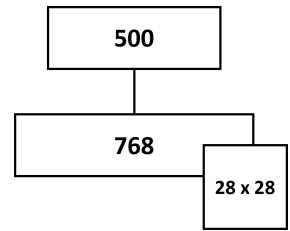
It's possible to highlight the following:

- The worse performance observed in the RBM with binary layers is due to the increased stochastic variance introduced by the Bernoulli sampling of hidden and visible unit
- The MNIST dataset presents lower training losses than Fashion-MNIST, which can be attributed to its simpler visual structure and less complex feature space
- The reconstruction performance improves as the number of training epochs increases
- The binary RBM exhibits significantly noisier reconstructions compared to the probability RBM, particularly when trained on the Fashion-MNIST dataset

(2 RBM) Impact of Sampling Iteration

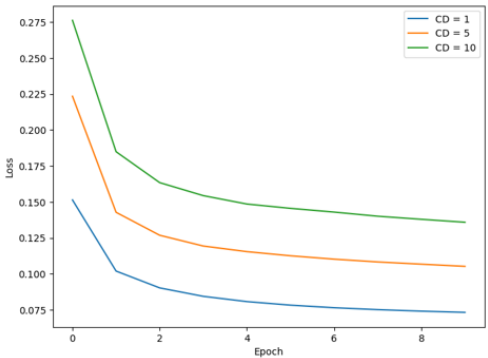
COPPE
UFRJ

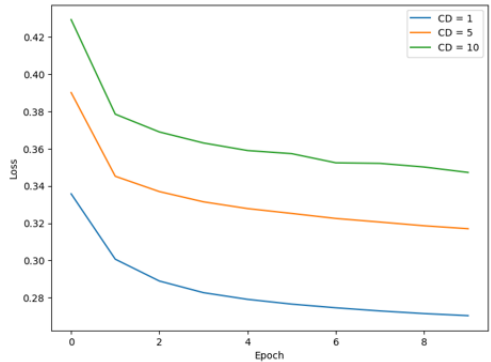
- **Objective**: Evaluate RBM with multiple iterations of Gibbs Sampling
- Binary cross entropy as loss function
- Datasets of MNIST and Fashion MNIST:
  - 60.000 (Train) and 10.000 (Test) instances



(2 RBM) MNIST and Fashion MNIST

COPPE
UFRJ





(2 RBM) Results

COPPE
UFRJ

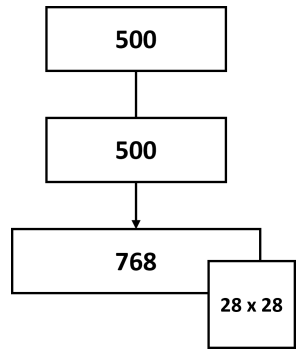
It's possible to highlight the following:

- The training performance becomes worse as the number of Gibbs sampling steps increases
- With more Gibbs sampling steps, the reconstructed samples move further away from the original data distribution - especially in the early stages of training when the weights are still random. Thus, the contrast between the data and model distributions becomes larger, leading to higher reconstruction error

(3 DBN) Analysis of Greedy Layer-Wise Training

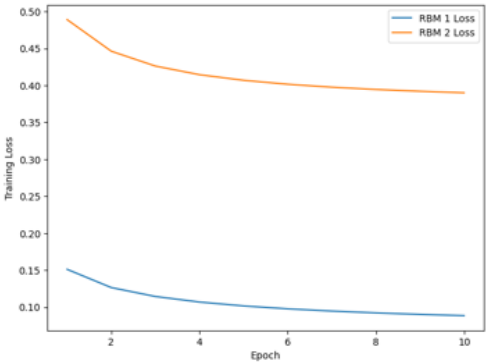
COPPE UFRJ

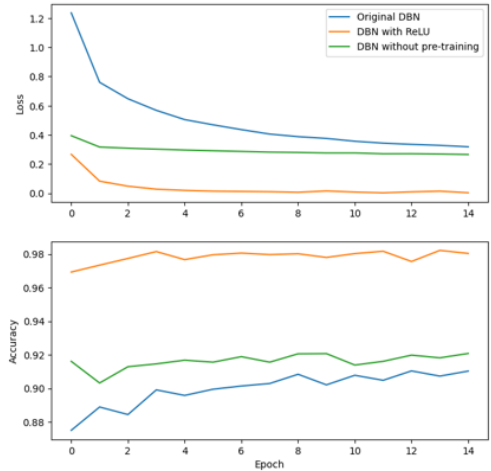
- **Objective**: Evaluate DBN's classification for three scenarios:
  - Original procedure (pre-training + fine-tuning)
  - Without pre-training
  - Original procedure with ReLU
- Model definitions:
  - Binary cross entropy for pre-training
  - Cross entropy loss for supervised training
  - Adam as optimizer
- Datasets of MNIST and Fashion MNIST:
  - 60.000 (Train) and 10.000 (Test) instances



(3 DBN) MNIST

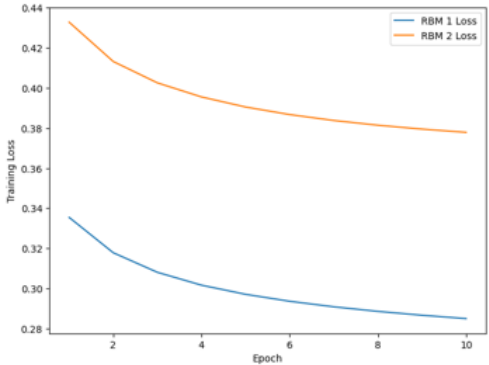
COPPE
UFRJ

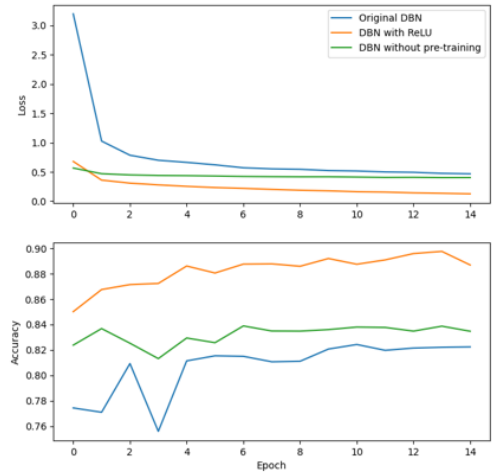




(3 DBN) Fashion MNIST

COPPE
UFRJ





(3 RBM) Results

COPPE
UFRJ

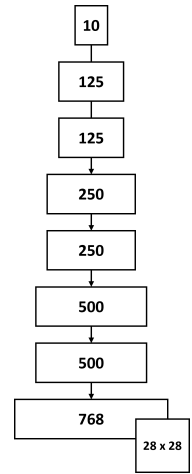
It's possible to highlight the following:

- The best DBN performance was achieved with the pre-training procedure combined with ReLU activation layers
- However, the DBN's performance improved without the pre-training procedure (green curve) compared to the original DBN (blue curve). This behavior is likely related to the use of the Adam optimizer, whose adaptive learning dynamics reduce the benefits of pre-training by already providing effective parameter updates and stabilization during training

(4 DBN) Evaluation of Multiple Layers

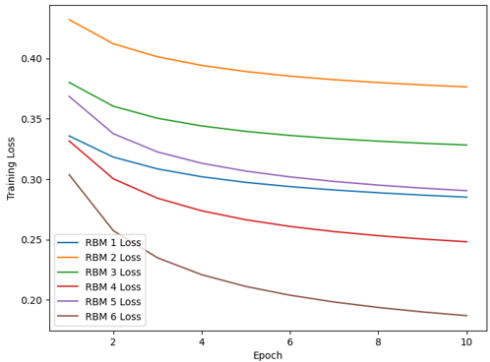
COPPE
UFRJ

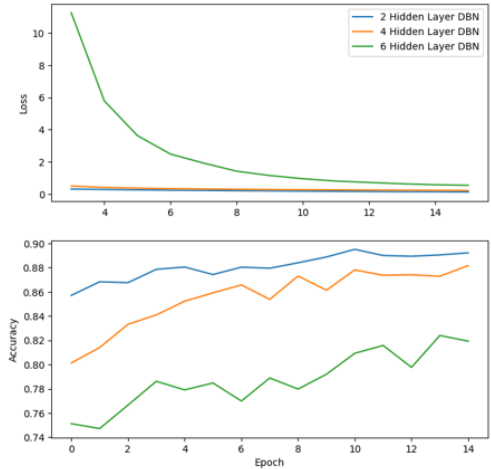
- **Objective**: Evaluate DBNs with multiple layers:
  - Input $\leftarrow 500 \leftrightarrow 500$
  - Input $\leftarrow 500 \leftarrow 500 \leftarrow 250 \leftrightarrow 250$
  - Input $\leftarrow 500 \leftarrow 500 \leftarrow 250 \leftarrow 250 \leftarrow 125 \leftrightarrow 125$
- Model definitions:
  - Binary cross entropy for pre-training
  - Cross entropy loss for supervised training
  - Adam as optimizer
- Fashion MNIST dataset:
  - 60.000 (Train) and 10.000 (Test) instances



(4 DBN) Fashion MNIST

COPPE
UFRJ





(4 RBM) Results

COPPE
UFRJ

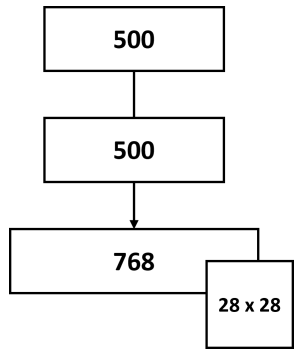
It's possible to highlight the following:

- The DBN's performance worsens as the number of hidden layers increases
- This degradation likely occurs because deeper architectures become harder to train effectively, leading to vanishing gradients and overfitting, especially when the amount of training data and regularization are limited

(5 DBN) Generation of Images

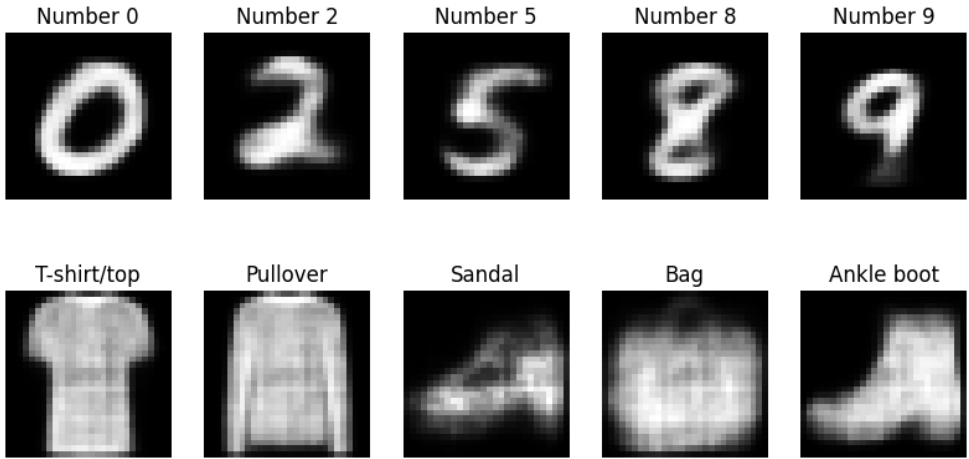
COPPE
UFRJ

- **Objective**: Use DBN to generate samples:
  - The samples were generated from the mean inference for each class
- Model definitions:
  - Binary cross entropy for pre-training
  - The supervised training procedure was not implemented
- Datasets of MNIST and Fashion MNIST:
  - 60.000 (Train) and 10.000 (Test) instances



(5 DBN) MNIST and Fashion MNIST

COPPE
UFRJ



(5 RBM) Results

COPPE
UFRJ

It's possible to highlight the following:

- The DBN could be successfully applied to data generation, even for the more complex Fashion-MNIST dataset
- This suggests that the model was able to capture meaningful representations of the input distribution, allowing it to synthesize realistic samples based on mean-valued latent representations

COPPE
UFRJ

Thank you for your attention

Guilherme Mota

Electrical Engineering Program
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
Universidade Federal do Rio de Janeiro

05/11/2025