Large-Scale Learning of Embeddings with Reconstruction Sampling

Yann N. Dauphin, Xavier Glorot, Yoshua Bengio

Université de Montréal

July 1, 2011

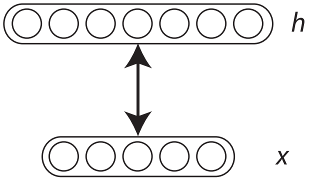
Background

- Surge of interest for unsupervised learning algorithms.
- (Hinton, 2006) shows that using the representation extracted by RBMs leads to superior classification..
- (Ranzato, 2006), (Bengio, 2006), et al. extend this results.

Motivation

- How to scale these algorithms to very sparse and very large input data?

Problem



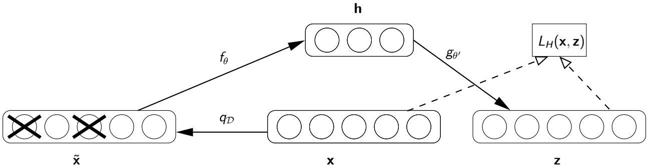
- Typically, 2 expensive mappings need to be computed.
    - Mapping from input to hidden representation.
    - Mapping from hidden to input representation.

Outline

Related work

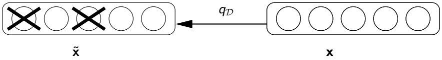
- (Morin and Bengio, 2005) propose a Tree-structured predictors.
- (Collobert and Weston, 2008) propose a ranking criterion estimated by Monte-Carlo sample.

Denoising Auto-Encoders



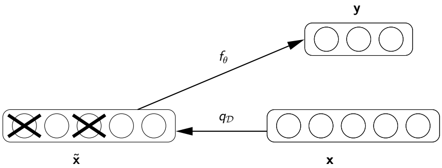
- Learning algorithm for unsupervised feature extraction (Vincent, 2008).

Denoising Auto-Encoders



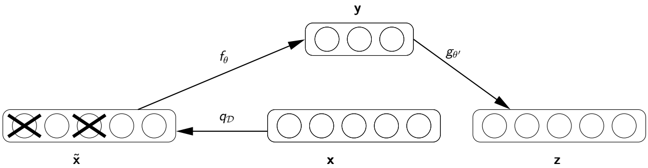
- $\mathbf{x} \in [0,1]^{d_x}$ is partially corrupted, yielding $\tilde{\mathbf{x}} \sim q_{\mathcal{D}}(\tilde{\mathbf{x}}|\mathbf{x})$.

Denoising Auto-Encoders



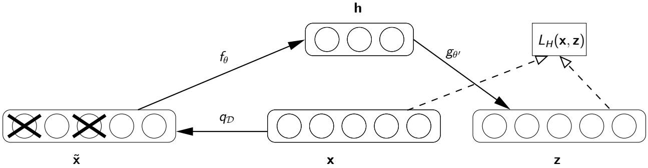
- Compute representation $\mathbf{h} = f_\theta(\tilde{\mathbf{x}}) = \mathrm{sigmoid}(\underbrace{\mathbf{W}^{(1)}}_{d_h \times d_x} \tilde{\mathbf{x}} + \underbrace{\mathbf{b}^{(1)}}_{d_h \times 1})$

Denoising Auto-Encoders



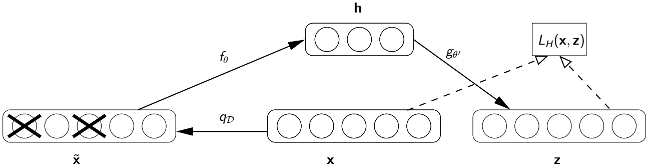
- Compute reconstruction $\mathbf{z} = g_\theta(\mathbf{h}) = \mathrm{sigmoid}(\underbrace{\mathbf{W}^{(2)}}_{d_x \times d_h}\mathbf{h} + \underbrace{\mathbf{b}^{(2)}}_{d_x \times 1})$

Denoising Auto-Encoders



- Train to minimize the cross-entropy $L(\mathbf{x}, \mathbf{z}) = \sum_k^d H(\mathbf{x}_k, \mathbf{z}_k)$.
- (Vincent, 2011) shows this is equivalent to training an EBM.

Denoising Auto-Encoders



- Train to minimize the cross-entropy $L(\mathbf{x}, \mathbf{z}) = \sum_k^d H(\mathbf{x}_k, \mathbf{z}_k)$.
- (Vincent, 2011) shows this is equivalent to training an EBM.

Sparse Dot

- Reduce encoder complexity with sparse dot: $f_\theta \in O(d_s \times d_h)$

Reconstruction Sampling

- Subsample the learning objective L.

$$\hat{L}(\mathbf{x}, \mathbf{z}) = \sum_{k}^{d} \frac{\hat{\mathbf{p}}_k}{\mathbf{q}_k} H(\mathbf{x}_k, \mathbf{z}_k)$$

- $\hat{\mathbf{p}} \in \{0, 1\}^{d_x}$ with $\hat{\mathbf{p}} \sim P(\hat{\mathbf{p}}|\mathbf{x})$ is the sampling pattern.
- $\mathbf{q}$ is are scalar weights, and if $\mathbf{q}_k = E[\hat{\mathbf{p}}_k | k, \mathbf{x}, \tilde{\mathbf{x}}]$ then objective is unbiased since $E[\frac{\hat{\mathbf{p}}_k}{\mathbf{q}_k} | k, \mathbf{x}, \tilde{\mathbf{x}}] = 1$.
- Reduces decoder complexity to $O(d_n \times d_h)$.

Sampling Distribution

- Minimize variance of the estimator.
- Sample bits where model will make an error.
- Let $\mathcal{C}(\mathbf{x}, \tilde{\mathbf{x}}) = \{k : \mathbf{x}_k = 1 \text{ or } \tilde{\mathbf{x}}_k = 1\}$, our heuristic:

$$P(\hat{\mathbf{p}}_k = 1|\mathbf{x}_k) = \begin{cases} 1 & \text{if } k \in \mathcal{C}(\mathbf{x}, \tilde{\mathbf{x}}) \\ |\mathcal{C}(\mathbf{x}, \tilde{\mathbf{x}})|/d_x & \textit{otherwise} \end{cases}$$

- Sample all 1s and equal amount of 0s.

Implementation

- The decoder is implemented as:

$$\mathbf{z} = \mathrm{sigmoid}(\texttt{SamplingDot}(\mathbf{h}, \mathbf{W}^{(2)}, \hat{\mathbf{p}}) + \mathbf{b}^{(2)})$$

Algorithm 1 $\texttt{SamplingDot}(\mathbf{A}, \mathbf{B}, \mathbf{C})$

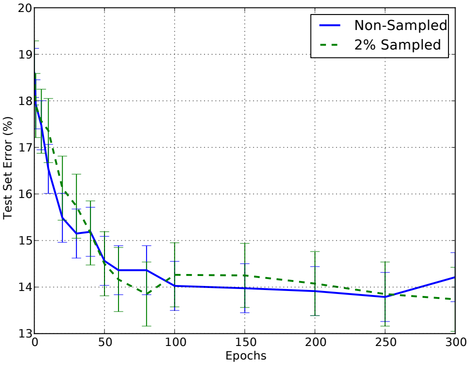
Input: $\mathbf{A} = [A_{ij}]_{M\times K}$, $\mathbf{B} = [B_{ij}]_{N\times K}$, $\mathbf{C} = [C_{ij}]_{M\times N}$
Output: $\mathbf{D} = [D_{ij}]_{M\times N}$

for $m = 1$ **to** M **do**
  for $n = 1$ **to** N **do**
    if $\mathbf{C}_{mn} \neq 0$ **then**
      $\mathbf{D}_m \leftarrow \texttt{DOT}(\mathbf{A}_m, \mathbf{B}_n)$
    end if
  end for

Datasets

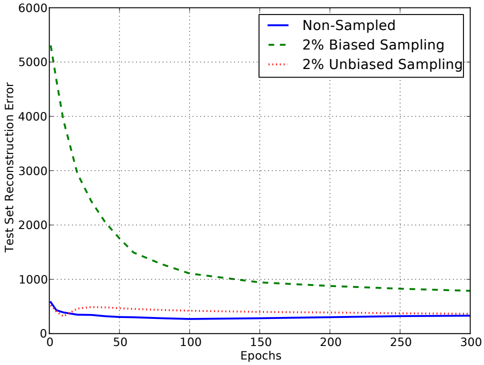
- **Amazon Multi-Domain Sentiment Dataset.** More than $340,000$ product reviews on 25 different domains.
- **Reuters Corpus Volume I** (RCV1-v2). Over 800,000 real-world news wire stories represented in bag-of-words vectors with 47,236 dimensions.

Convergence



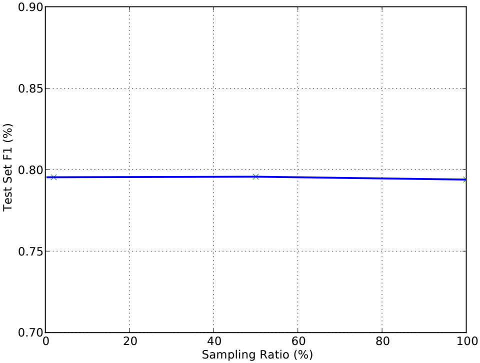
Amazon

Bias


Non-Sampled
2% Biased Sampling
2% Unbiased Sampling
Test Set Reconstruction Error
Epochs
0
1000
2000
3000
4000
5000
6000
50
100
150
200
250
300

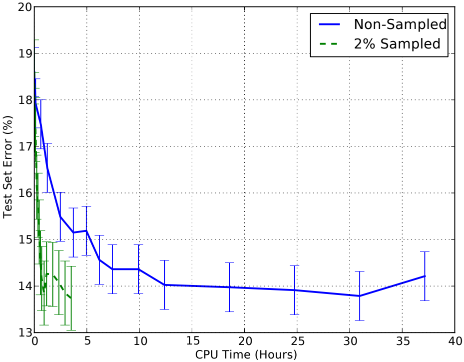
Amazon

Quality of the representation



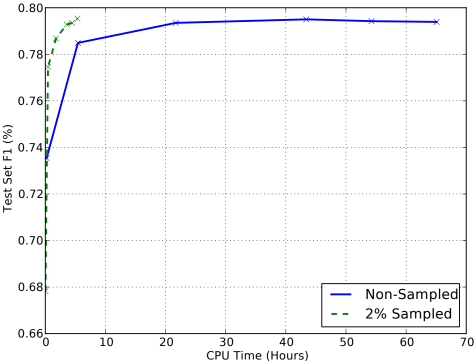
RCV1-v2

Speed-ups



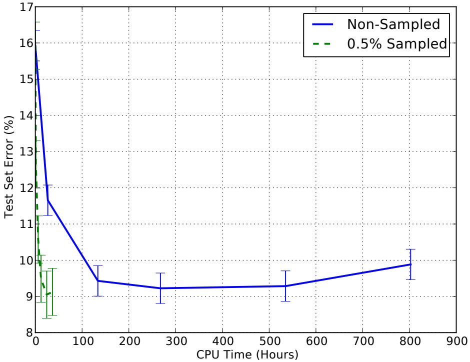
Amazon

Speed-ups



RCV1-v2

Speed-ups



Full Amazon

Conclusion

- Introduced simple speed-up technique.
- Unbiased estimator.
- Same quality of representation.
- Speed-ups up to 20x.