Proyecto: Binary Variational Semantic Hashing

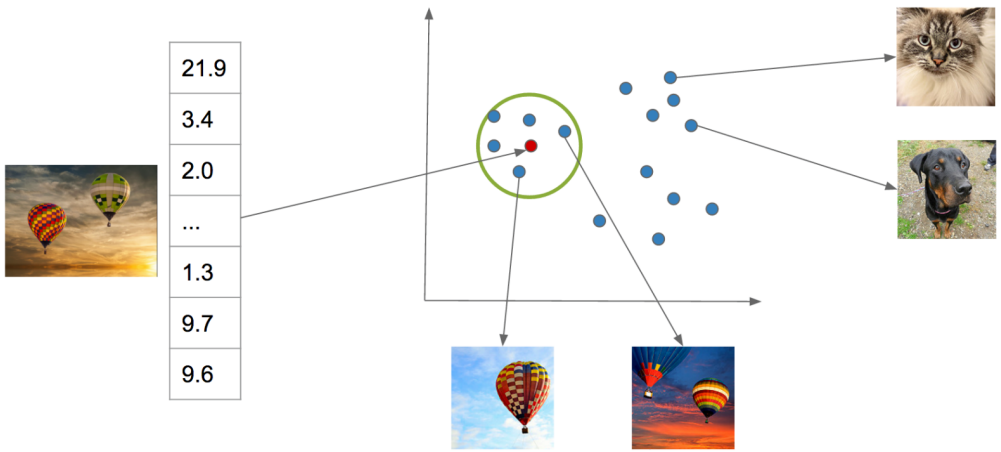
Francisco Mena

UTFSM - Departamento de Informática

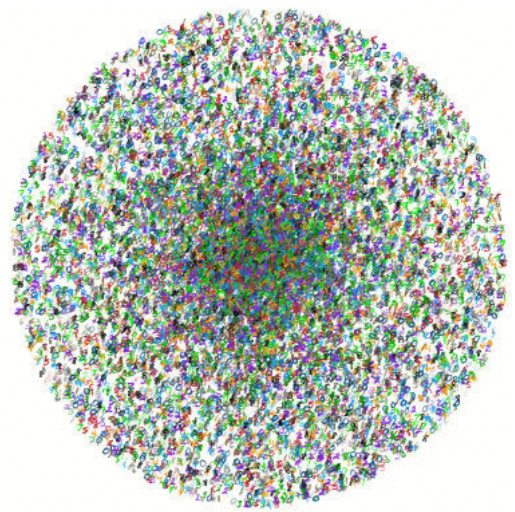
27 de Noviembre

Motivación

- Búsqueda de contenido relevante en una colección gigante de datos puede ser bastante costoso



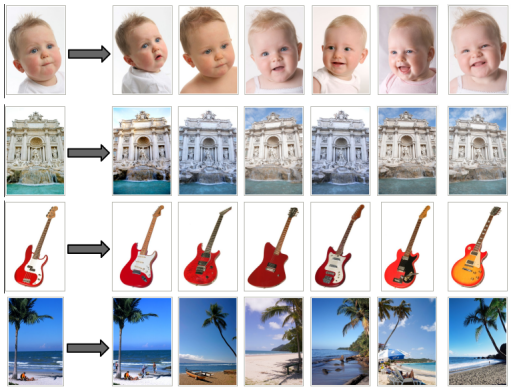
- Métodos tradicionales computan similaridad en el espacio original (BOW,TF-IDF) o en *vector space*, lo cual no escala computacionalmente



Problema

Problema: *similarity search*/*proximity search*

- Encontrar/Recuperar objetos similares dada un objeto como consulta
- También conocido como *content-based retrieval*



Solución: *Hashing* (*hash-based similarity search*)

- Acelera la búsqueda de objetos similares
- Asigna códigos binarios compactos (baja dimensionalidad) a cada objeto
- Propiedad Semántica: Objetos semánticamente similares generan códigos similare

Trabajos previos

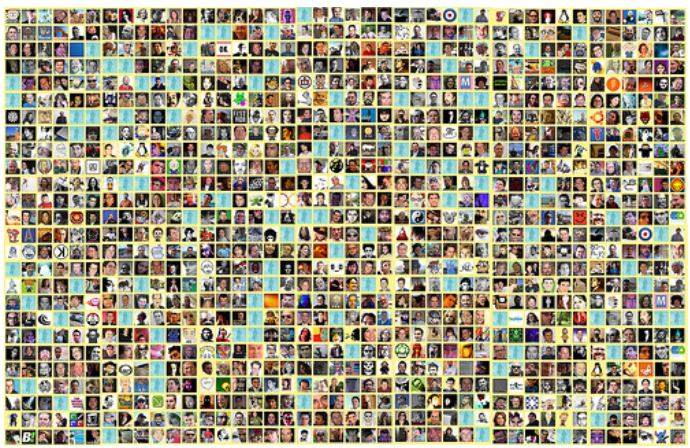
Aprendizaje de función de *hashing* de manera no supervisada:

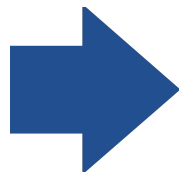
- Semantic Hashing (Salakhutdinov y Hinton, 2007)
  1. RBM's para generar distribución de variables latentes binarias
  2. Representación de documentos como *word-count* (TF)
- Spectral Hashing (Weiss et al. 2009)
  1. Problema de grafo, solución asemeja a spectral clustering
  2. Binariza los vectores propios con un *treshold* de cero
- Variational Deep Semantic Hashing (Chaidaroon, 2017)
  1. Utiliza VAE's para generar una distribución sobre variable latentes continuas
  2. Binariza tomando como *treshold* la mediana

Propuesto

Objetivo

Mejorar el aprendizaje no supervisado de *hashing* semántico a través de un modelo probabilista que se adapte correctamente a la necesidad de generar códigos binarios





$$\begin{matrix} \cdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \cdots \end{matrix}$$

Métodos y Metodologías

¿Qué?

- *Hashing* semántico con redes neuronales utilizando Variational Autoencoder (VAE) para inferencia probabilísta de variable latente binaria

¿Técnicas?

- Representación de palabras como *word count* (TF) o binario
- Redes *feed forward*
- Variable latentes **discretas** en VAE
  - Gracias al truco de reparametrización Gumbel-Max suavizado

Modelo

Formulación

$$\ell(\vec{d}) = \mathbb{E}_Q \left[\sum_i^N \log P(w_i \mid \vec{b}; \theta) \right] - D_{KL} \left(Q(\vec{b} \mid \vec{d}; \phi) \mid\mid P(\vec{b}) \right) \quad (1)$$

- $\vec{d}$: representación vectorial de documento (BOW)
- w_i: representación one-hot de la palabra (TF, binario)
- $\vec{b}$: código binario
- Q: *encoder*, P: *decoder*

Se asume que $Q(\vec{b} \mid \vec{d}; \phi)$ aproxima una distribución Bernoulli por cada componente (función sigmoidal)

Modelo Neuronal

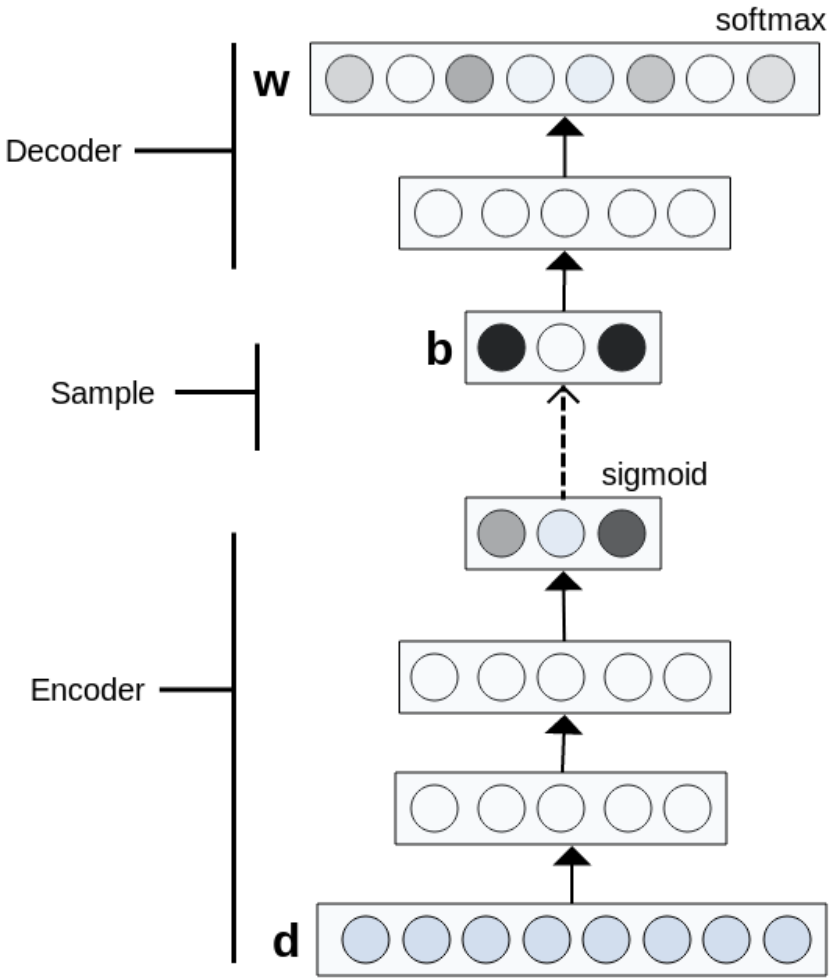


Figure 1: modelo *Binary Variational Semantic Hashing*

Truco de reparametrización

Formulación

- $\vec{b}_d = (b_d^{(1)}, b_d^{(2)}, \ldots, b_d^{(K)})$
- Truco Gumbel-Softmax (caso binario), con $U \sim \text{Uniform}(0, 1)$

$$b_d^{(k)} = \sigma\left(\left(\log \frac{Q(b^{(k)} \mid \vec{d})}{1 - Q(b^{(k)} \mid \vec{d})} + \log \frac{U}{1-U}\right)/\lambda\right), \quad \forall k$$

$$\mathcal{L} = \sum_{\vec{d} \in D} D_{KL}\left(Q(\vec{b} \mid \vec{d}) \,||\, P(\vec{b})\right) - \left(\sum_{w \in \vec{d}} w \cdot \log P(w \mid \vec{b}_d)\right) \quad (2)$$

Forma de validación

- Métricas: *precision* y *recall* sobre conjunto de pruebas
  - Se recuperan objetos en base a distancia de hamming
- Variación de métricas vs el radio de búsqueda
- Curvas *precision* y *recall* variando radio de búsqueda
- *Baseline*: Variatonal Deep Semantic Hashing (VDSH)
- ¿Dónde?

Dataset	**Documentos**	**Clases**
20Newsgroup	18.828	20
Reuters Corpus I	800.000	103
Reuters Corpus II	804.414	103
Reuters21578	10.788	90
TMC	28.515	22
SearchSnippets	12.000	8

Proyecto: Binary Variational Semantic Hashing

Francisco Mena

UTFSM - Departamento de Informática

27 de Noviembre