A Binary Variational Autoencoder for Hashing

CIARP 2019

Francisco Mena and Ricardo Ñanculef

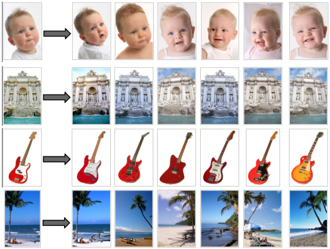
Departamento de Informática
Universidad Técnica Federico Santa María
Chile

October 2019

Summary

Big Data Challenges

- In recent years, our capacity to generate and collect data has growth explosively. Most of this content arrives in the form of multimedia data : text, images and video.
  - **Google :** ≥ 200 billion emails every day.
  - **Instagram :** ≥ 100 million video and photos uploaded every day.
  - **Large Synoptic Survey Telescope (LSST) :** ≥ 30 Terabytes of image data every night.
- The size of these datasets challenges traditional ways to approach many computational problems, even "simple" ones such as **similarity search**.



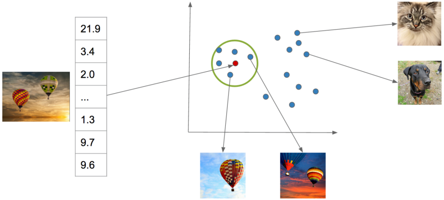
Target Task : *similarity search*

- Find/Retrieve elements in a database that are similar to a sample (query) object.
- A.k.a. *content-based retrieval*.

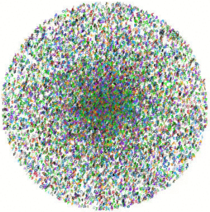
EX UMBRA IN SOLEM

Similarity Search

- Classic approach : Compute a “descriptor” of the database items and queries ; then retrieve items which are “similar” to a given query.



- Traditional methods for text compute similarity in the original word-count space (e.g. TF-IDF). This is slow for large vocabularies and does not capture semantic similarity between texts.
- Continuous/dense descriptors are difficult to store and index efficiently as classic data structures fail in high dimensions (> 10 dims).

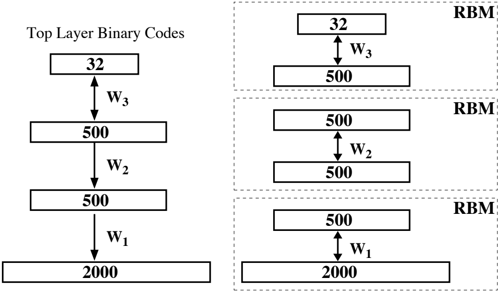


Contribution

- We propose to learn hash codes $h(x)$ by using a **binary variational auto-encoder (BAE)**, a deep generative model with a Bernoulli latent layer that can be trained with standard back-propagation.
- Our method improves on previous methods by addressing the **quantization loss** $\|h(x) - \phi(x)\|$ introduced by formulations based on thresholding a continuous embedding $\phi(x)$.
- Our method is more interpretable and produces balanced hash tables thanks to the use of non-informative priors.
- For space constraints we focus on text retrieval applications and the **unsupervised** case, but extensions are possible.

Seminal Work

- Semantic Hashing (Salakhutdinov & Hinton, 2008)
  - Deep stack of RBM to generate a binary latent variables.
  - Modern perspective : An stochastic auto-encoder where the encoder is obtained by reversing the arcs of the decoder.
  - Difficult to train : it is based on layer-wise pre-training.



The Shallow Menace

Perhaps, for efficiency reasons, many sub-sequent works focused on shallow architectures.

- Spectral Hashing (Weiss et al. 2009)
  - Problem of partition a graph, related to spectral clustering.
  - Threshold values with zero.
- Kernelized Random projections (ref)
  - Project data onto random hyper-planes defined in a kernel-induced feature space.
- PCA-based hashing (ref)
  - Project data on PCA directions and then minimize the quantization error.

Depth Strikes Back

More recent works restored depth, but kept stochasticity away, to allow the use of gradient-based optimization.

- (2, ref)
  - Shallow autoencoder (AE) that minimizes the reconstruction error with a binary constrain.
- (9, ref)
  - Deep PCA minimizing variance and quantization loss.
- (4, ref)
  - deep PCA + linear decoder (asymmetric autoencoder).
  - minimizes reconstruction error with binary constrain similar to 2

Return of the VAE

Recently, it has been shown that deep generative models, in particular **Variational Autoencoders (VAEs)**, can be succesfully used for topic modeling and text hashing.

- Variational Deep Semantic Hashing (VDSH, ref)
  - Learn a continous VAE, that is a stochastic auto-encoder with a gaussian latent representation and prior.
  - Threshold continous latent variable around median/zero
- (12, ref)
  - Learn a categorical VAE for discover topics in text documents.

Proposal Focus

Focus :

- Learn a hash function $h(\cdot)$ using a **deep probabilistic graphical model**
- Improve the results on similarity search on the unsupervised case

How :

- Using the VAE framework (Kingma et al. 2013)
- Model explicitly a binary latent variable $b \in \{0, 1\}^B$

Advantages :

- Interpretability of learned representation
- Reduce the error introduce in the quantization step

Model Architecture

Encoder :

- Codify input pattern into a multi-variate Bernoulli distribution
- $q_\phi(b|x) = \text{Ber}(\alpha(x))$
- $\alpha(x) = p(b = 1|x)$ is modeled using a neural net $f(x; \phi)$

Decoder :

- Reconstruct input pattern from the binary codes : $p_\theta(x|b)$
- For text representation, a multinomial distribution of tokens : $p(x|b) = \prod_{w \in x} p(w|b)^{tf_w}$
- $p(w|b)$ is modeled using a neural net $g(b; \theta)$

As VAE framework, the learning goal is based on lower bound of $\ell(\theta, \phi; D)$:

$$\ell(\theta, \phi; x^{(\ell)}) \geq \mathcal{L} = \mathbb{E}_{q_\phi(b|x^{(\ell)})}\left[\log p_\theta(x^{(\ell)}|b)\right] - D_{\text{KL}}\left(q_\phi(b|x^{(\ell)})||p_\theta(b)\right) , \quad (1)$$

Optimization details

Re-parameterization via Gumbel-Softmax

$b_{i,\ell} \sim$ Ber $\left(\alpha_i(x^{(\ell)})\right)$, $\epsilon_i \sim \mathcal{U}(0,1)\ \ \forall i \in [B]$

$$\hat{b}_{i,\ell} = \sigma\left(\left(\log \frac{\alpha_i(x^{(\ell)})}{1-\alpha_i(x^{(\ell)})} + \log \frac{\epsilon_i}{1-\epsilon_i}\right)/\lambda\right) \tag{2}$$

$\hat{b}_{i,\ell}$ converges to $b_{i,\ell}$ in the sense that $P(\lim_{\lambda \to 0} \hat{b}_{i,\ell} = 1) = \alpha_i(x)$

We set the *temperature* $\lambda = 2/3$ as previous experiments show

Priors

We model $p_\theta(b_i) =$ Ber$(0.5)\ \ \forall i \in [B]$. The KL divergence is expressed as

$$D_{\text{KL}}\left(q_\phi(b|x^{(\ell)})||p_\theta(b)\right) = B\log 2 + \sum_i^B \alpha_i(x)\cdot\log\alpha_i(x) + (1-\alpha_i(x))\cdot\log\left(1-\alpha_i(x)\right) \tag{3}$$

Implementation

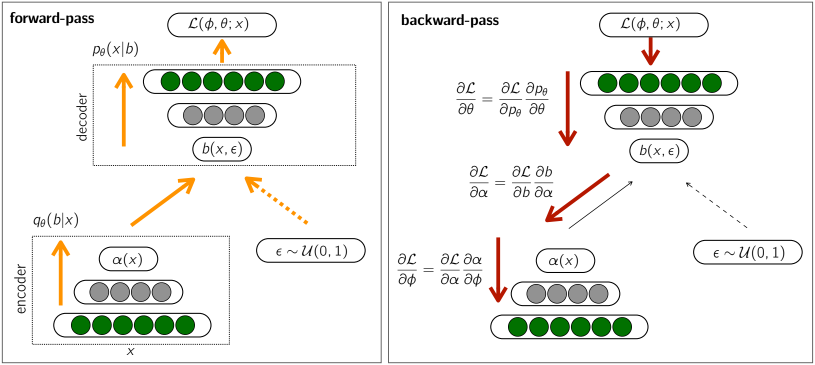


FIGURE – Illustration of the forward (*orange*) and backward (*red*) pass implementing the proposed method as a deep neural net. The dashed line represents a stochastic layer.

Quantization

- Discretization is no required (into sampled step)
- Deterministic codes : threshold is done on the probability encoded

$$b = 1(\alpha(x) - \frac{1}{2}) \quad (4)$$

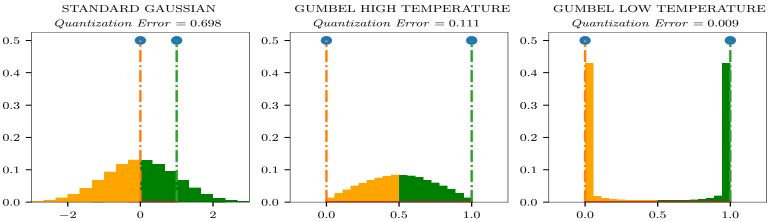


FIGURE – 1-bit quantization of a Gaussian variable (standard VAE) and two Gumbel-Softmax variables (B-VAE) at different temperatures. In practice, all the yellow/green points are rounded to $0/1$ to obtain binary codes. A Gumbel-Softmax distribution at low temperature reduces the quantization error inducing a saturation around 0/1.

Baseline and Dataset

- The baseline is VDSH (continuous VAE for hashing) as improve result on previous unsupervised techniques
- Based on early experiments on neural net architecture we define :
  - VDSH architecture of original paper
  - B-VAE symmetric architecture of VDS
- We evaluate our method on text retrieval tasks
  - **20 Newsgroups** : 18000 long documents, 20 mutually exclusive classes
  - **Reuters21578** : 11000 news documents, 90 non-exclusive tags (topics)
  - **Google Search Snippets** : 12000 short documents, 84 mutually exclusive domains

Representation

Pre-processing

1. Lower-case
2. Remove extra-space, stop-words and any character that is not a letter
3. Lemmatize step
4. Remove lemmas of length smaller than 3.

We use TF representation tf_d of each document (10^4 most frequent lemmas)

- Early experiments reveled that sub-linear TF make training more stable : $\log(\text{tf}_\text{d} + 1)$

Evaluation

- Hash table is build embeding the training set (with the trained models)
- Each test or validation document are provided to the system as a query and used to retrieve similar documents on Hash table
- Two items were considered **similar** if they have at least one label in common
- We consider two **querying** methods : *top*-K and *radius search*
- The results are evaluated using precision (P) and recall (R)

Results

tablas

Conclusion

- We have investigated the use of a variational autoencoder with binary latent variables to learn hash codes
  - Easier to interpret
  - Reduces the quantization error of thresholding continuous codes
  - Consents the use of back-propagation for training.
- Experiments on unsupervised text hashing show that the method is more effective for IR than its continuous counterpart
- As **future work**, we plan to evaluate the model on image retrieval tasks using convolutional nets and to handle semi-supervised scenarios

A Binary Variational Autoencoder for Hashing

CIARP 2019

Francisco Mena and Ricardo Ñanculef

Departamento de Informática
Universidad Técnica Federico Santa María
Chile

October 2019