Block Coordinate Descent Method for Cross-Modal Retrieval

Zihao Wang[1]

[1] College of Information Sciences and Technology
Penn State University

4 Dec 2018

Cross-modal Retrieval

Modal : Datatype

View : Each type of data is treated as a single view

Cross Modal : Returns relevant results of one modality in response to a query of another modality

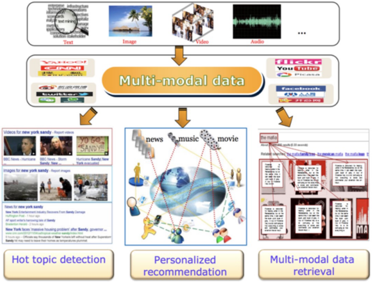

FIGURE – Multi-Modal data

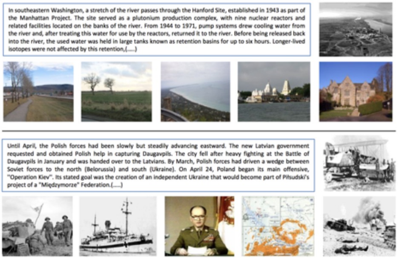

FIGURE – Cross-modal Retrieval

Definition and Problem Description

Training Data : $\mathcal{D} = \{((\mathbf{I}_1, \mathbf{T}_1), y_1), \ldots, ((\mathbf{I}_N, \mathbf{T}_N), y_N)\}$, imagine and text pair

y_i : label

N : sample size

Image Data : $\mathbf{I}_i = (\mathbf{I}_i^{(1)}, \ldots, \mathbf{I}_i^{(V_1)}), \mathbf{I}_i^{(\nu_1)} \in \mathbb{R}^{d_{\nu_1}}$

$\mathbf{I}_i^{(\nu_1)}$: the image feature vector in the view ν_1

d_{ν_1} : the dimensionality of the view ν_1

Text Data : $\mathbf{T}_i = (\mathbf{T}_i^{(1)}, \ldots, \mathbf{T}_i^{(V_2)}), \mathbf{T}_i^{(\nu_2)} \in \mathbb{R}^{d_{\nu_2}}$

$\mathbf{T}_i^{(\nu_2)}$: the text feature vector in the view ν_2

d_{ν_2} : the dimensionality of the view ν_2

Purpose : Build a function $f(\{(\mathbf{I}_i, \mathbf{T}_i)\}) : \mathcal{I} \to \mathcal{T}$ or $f(\{(\mathbf{I}_i, \mathbf{T}_i)\}) : \mathcal{T} \to \mathcal{I}$

Approach : Learn a discriminant function $F : \mathcal{I} \times \mathcal{T} \to \mathbb{R}$ to predict the optimal output $\mathbf{T}^*$

$$\mathbf{T}^* = f(\mathbf{I}; \mathbf{w}) = \arg\max_{\mathbf{T} \in \mathcal{T}} F(\mathbf{I}, \mathbf{T}; \mathbf{w}) \tag{1}$$

$\mathbf{w}$: the parameter needed to be learned

Empirical Risk

With the training data coming from P topics $\{\mathcal{D}_{p=1}^{P}\}$, we can write the regularized **empirical risk** $\mathcal{R}$ for cross-modal retrieval problem as :

$$\mathcal{R}(\{F_p\}_{p=1}^{P}) = \sum_{p=1}^{P}(\mathcal{L}_p(F_p(\mathbf{I}_p.\mathbf{T}_p; \mathbf{w}_p), y) + \lambda\Omega(\mathbf{w}_p)) \tag{2}$$

where $\Omega(\cdot)$ is a regularization term, and $\lambda > 0$ is the regularization hyper-parameter. The empirical loss of the training data from each topic $\mathcal{L}_p$ is

$$\mathcal{L}_p\left(F_p\left(\mathbf{I}_p, \mathbf{T}_p; \mathbf{w}_p\right), \mathbf{y}\right) = \sum_{i=1}^{N_p} \frac{1}{N_p} \ell\left(F_p(\mathbf{I}_{p,i}, \mathbf{T}_{p,i}; \mathbf{w}_p), y_{p,i}\right) \tag{3}$$

where ℓ is the prescribed loss function, here I use squared loss, and N_p is the sample number of the topic p

Multi-view Data Fusion through Tensor Modeling

Multi-view Data Fusion :

$$f\left(\left\{x^{(v)}\right\}_{v=1}^{V}\right) = \sum_{S=1}^{P} \sum_{i_1=0}^{d_1} \cdots \sum_{i_V=0}^{d_1} w_{i_1,\cdots,i_V,s}\left(\Pi_{v=1}^{V} \mathbf{z}_{i_\nu}^{(\nu)}\right) \tag{4}$$

$\mathbf{x}^{(v)} \in \mathbb{R}^{d_v}$, the input multi-view data

$\mathbf{z}^{(\nu)} = [1; \mathbf{x}^{(\nu)}]$

$\{w_{i_1}, \cdots, i_V, s\}$ the weight tensor to be learned, can be factorized into R factors as $[\![\Theta^{(1)}, \ldots, \Theta^{(V)}]\!]$

$\Theta^{(\nu)} \in \mathbb{R}^{(d_v+1)\times R}$, the shared structure matrix for the v-th view

After CP factorization :

$$f\left(\left\{x^{(\nu)}\right\}_{\nu=1}^{V}\right) = \Pi_{\nu=1}^{V} \odot \left(z^{(\nu)T}\Theta^{(\nu)}\right)^{T} \tag{5}$$

Joint Optimization Problem

The joint optimization problem following the regularization formulation :

$$\min \mathcal{R}(\{\Theta^{(v)}\}) = \mathcal{L}_p(f(\{\mathbf{x}_I^{(v_1)}\}, \{\mathbf{x}_T^{(v_2)}\}), \mathbf{y}) + \lambda\Omega_\lambda(\{\Theta_I^{(v)}\}, \{\Theta_T^{(v)}\}) \tag{6}$$

y the label

$\mathcal{L}$ the loss function

$\{\Theta_I^{(v)}\}, \{\Theta_T^{(v)}\}$ can be obtained by solving the problem

Ω_λ the regularization terms, I use Frobenius norm

Though all parameters are convex, together Eqn(6) is non-convex with all the parameters.

Framework of Block Coordinate Descent

Algorithm 1 Block coordinate descent

Initialization: choose $(\mathbf{x}_1^0, \cdots, \mathbf{x}_s^0)$
for $k = 1, 2, \cdots$ **do**
  for $i = 1, 2, \cdots, s$ **do**
    update $\mathbf{x}_i^k$ with all other blocks fixed
  end for
  if stopping criterion is satisfied **then**
    return $(\mathbf{x}_1^k, \cdots, \mathbf{x}_s^k)$.
  end if
end for

Throughout iterations, each block x_i is updated by one of the three update schemes :

1. **Block minimization**
2. Block proximal descent
3. Block proximal linear

Alternating Block Coordinate Descent

STEP 1 : Fix $\boldsymbol{\alpha}$, and $\boldsymbol{\Theta}_T^{(v_2)}$, minimize $\boldsymbol{\Theta}_I^{(v_1)}$

$$\frac{\partial \mathcal{L}_p}{\partial f_p} \frac{\partial f_p}{\partial \boldsymbol{\Theta}_I^{(v_1)}} = \alpha_p \mathbf{Z}_{p,I}^{(v_1)}((\frac{\partial \mathcal{L}_p}{\partial f_p}) * \boldsymbol{\Pi}_{p,I}^{(-v_1)}) \tag{7}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_p}{\partial f_p} &= \frac{1}{N_p}[\frac{\partial \ell_{p,1}}{f_{p,1}}, \cdots, \frac{\partial \ell_{p,N_p}}{f_{p,N_p}}]^T \in \mathbb{R}^{N_p} \\
\boldsymbol{\Pi}_{p,I}^{(-v_1)} &= [\boldsymbol{\pi}_{I,1}^{(-v_1)}, \cdots, \boldsymbol{\pi}_{I,N_p}^{(-v_1)}]^T \\
\boldsymbol{\pi}_I^{(-v_1)} &= \Pi_{v_1'=1, v_1' \neq v_1}^{V_1} * (\mathbf{z}_I^{(v_1')T} \boldsymbol{\Theta}_I^{(v_1')})^T \in \mathbb{R}^R
\end{aligned}$$

Similarly, for $\boldsymbol{\Theta}_T^{(v_2)}$

$$\frac{\partial \mathcal{L}_p}{\partial f_p} \frac{\partial f_p}{\partial \boldsymbol{\Theta}_T^{(v_2)}} = (1-\alpha_p) \mathbf{Z}_{p,T}^{(v_2)}((\frac{\partial \mathcal{L}_p}{\partial f_p}) * \boldsymbol{\Pi}_{p,T}^{(-v_2)}) \tag{8}$$

$$\boldsymbol{\Pi}_{p,T}^{(-v_2)} = [\boldsymbol{\pi}_{T,1}^{(-v_2)}, \cdots, \boldsymbol{\pi}_{T,N_p}^{(-v_2)}]^T$$

STEP 2 : Update α

$$\frac{\partial \mathcal{R}}{\partial \boldsymbol{\alpha}} = [(\frac{\partial \mathcal{L}_1}{\partial f_1})^T \boldsymbol{\Delta}_1, \cdots, (\frac{\partial \mathcal{L}}{\partial f})^T \boldsymbol{\Delta})] \tag{9}$$

$$\boldsymbol{\Delta}_p = \mathbf{f}_{p,I} - \mathbf{f}_{p,T}, \forall p \in [1:P], \boldsymbol{\Delta}_p \in \mathbb{R}^{N_p}$$

Experiments

Dataset : The NUS-WIDE dataset is a real-world image dataset created by Lab for Media Search in National University of Singapore. This dataset contains 81 topics.

TABLE – Imagine to Text Retrieval

Modle	mAP	Precision
JRL	0.5432	0.5010
SMFH	0.5974	0.4658
TM	0.7011	0.7467

TABLE – Text to Imagine Retrieval

Modle	mAP	Precision
JRL	0.5199	0.5218
SMFH	0.5596	0.4973
TM	0.7693	0.7748

mAP : the mean of the average precision scores for each query

$$mAP = \frac{\sum_{q=1}^{Q} \mathrm{AveP}(q)}{Q} \tag{10}$$

Q : the number of queries