Online Asymmetric Similarity Learning for Cross-Modal Retrieval

Yiling Wu[1,2], Shuhui Wang[1], Qingming Huang[1,2]
[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

Motivations:

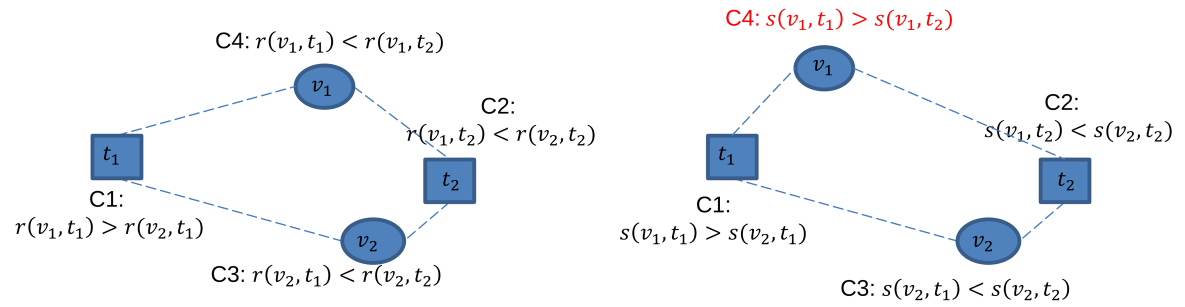
- The critical problem in cross-modal retrieval task is how to measure the similarity between data from different modalities.
- The relations between images and texts are highly asymmetric. There are two kinds of relative similarities that can be used.
- CNN features are state-of-the-art features, but there are many CNN layers. Choosing which layer to use is a difficult problem.

Contributions:

- We propose an **online learning method** to learn the similarity function between heterogeneous modalities by preserving the bi-directional relative similarity in the training data.
- We extend it to an **online multiple kernel learning method** to address the problem of combining different layers of CNN features for cross-modal retrieval.

Learning Bi-direction Relative Similarity:

- Consider learning a bilinear similarity function $s(v_i, t_j) = {v_i}^T \mathbf{W} t_j$
- Bi-directional relative similarity constraints are indispensable for modeling the cross-modal relation.



- We expect the similarity function to satisfy the following two conditions simultaneously:

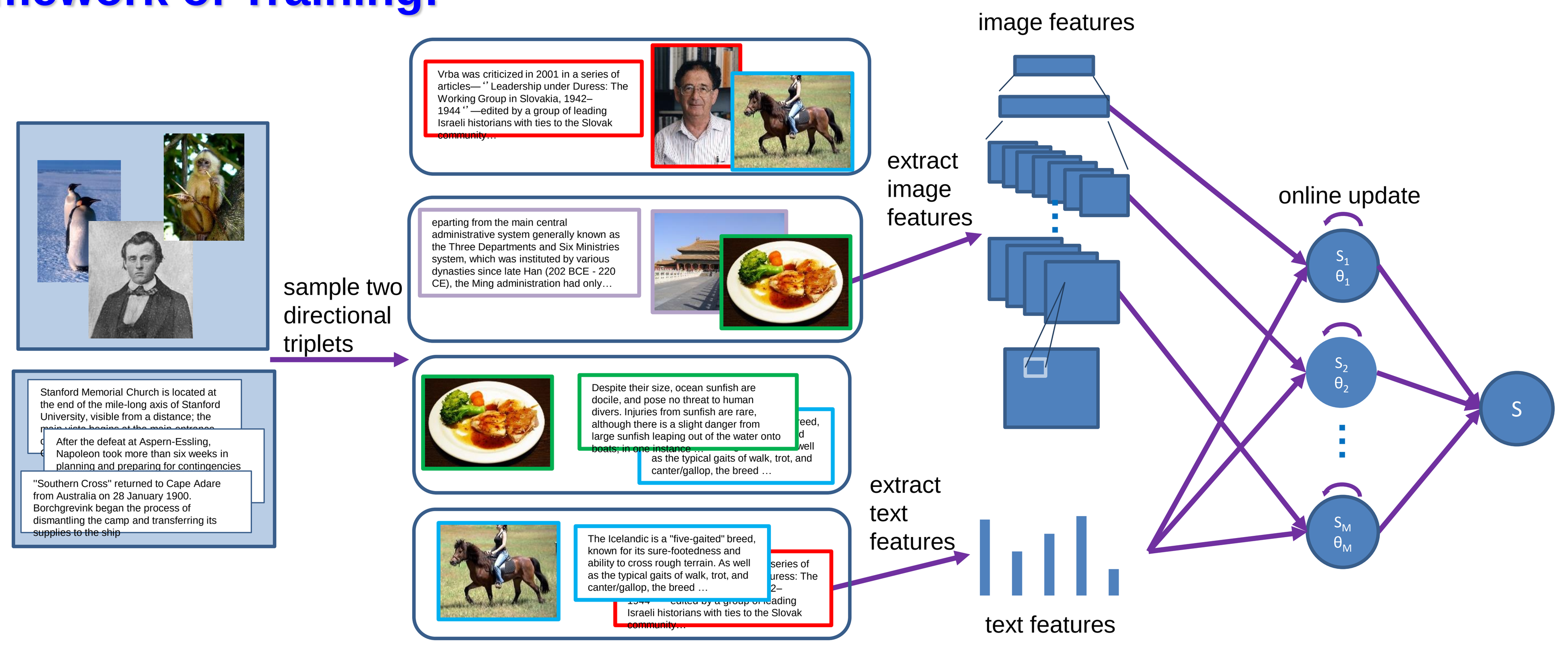
$$l_v(\mathbf{W}; v_i, t_i^+, t_i^-) = \max\{0, s(v_i, t_i^-) - s(v_i, t_i^+) + 1\}$$

$$l_t(\mathbf{W}; t_j, v_j^+, v_j^-) = \max\{0, s(v_j^-, t_j) - s(v_j^+, t_j) + 1\}$$

$$L(\mathbf{W}; D_{train}) = \sum_{\pi_i \in \Pi^v} l_v(\mathbf{W}; v_i, t_i^+, t_i^-) + \sum_{\pi_i \in \Pi^t} l_t(\mathbf{W}; t_i, v_i^+, v_i^-)$$

- The model is trained by the Passive-aggressive (PA) algorithm. We call this method Cross-Modal Online Similarity function learning (CMOS).

Framework of Training:



Online Multiple Kernel Learning:

- To select multiple CNN layers, we derive its multiple kernel extension which called Cross-Modal Online Multiple Kernel Similarity function learning (CMOMKS).
- We extend the primal model to its dual from. A pair of kernel is used for each similarity function:

$$s(v,t) = \sum_{\pi_i \in \Pi^v} \tau_i k^v(v, v_j)(k^t(t_j^+, t) - k^t(t_j^-, t)) + \sum_{\pi_i \in \Pi^t} \tau_i (k^v(v, v_j^+) - k^v(v, v_j^-)) k^t(t_j, t)$$

- We want to learn the coefficients of linear combinations of the M pairs of kernels, while at the same time we learn each similarity function. Let $f(v,t) = \sum_{j=1}^{M} \theta_j s_j(v,t)$, we consider:

$$\min_{\theta \in \Delta} \min_{\{s_j\}_{j=1}^{M}} \frac{1}{2} \sum_{i=1}^{M} \|L_i\|_{HS}^2 + C\left(\sum_{\pi_i \in \Pi^v} l_v(f; \pi_i) + \sum_{\pi_i \in \Pi^t} l_t(f; \pi_i)\right)$$

- At every iteration, for each of the M pairs of kernels, e.g., $(k_j^v; k_j^t)$, we apply the PA algorithm to find the optimal coefficient for the kernelized similarity function, and then apply the Hedging algorithm to update the combination weight by

$$\theta_j(i) = \theta_j(i-1)\beta^{z_j(i)}$$

where $z_j(i)$ equals to 1 when $s(v_i; t_i^+) - s(v_i; t_i^-) \le 0$ or $s(v_i^+; t_i) - s(v_i^-; t_i) \le 0$, and 0 otherwise.

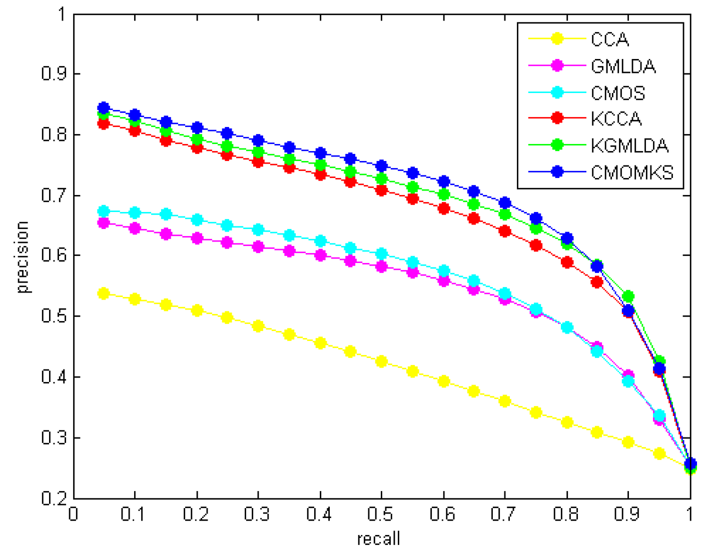
- Mistake bound can be easily obtained according to the mistake bounds of the PA and the Hedge algorithm.

Experiment Setup:

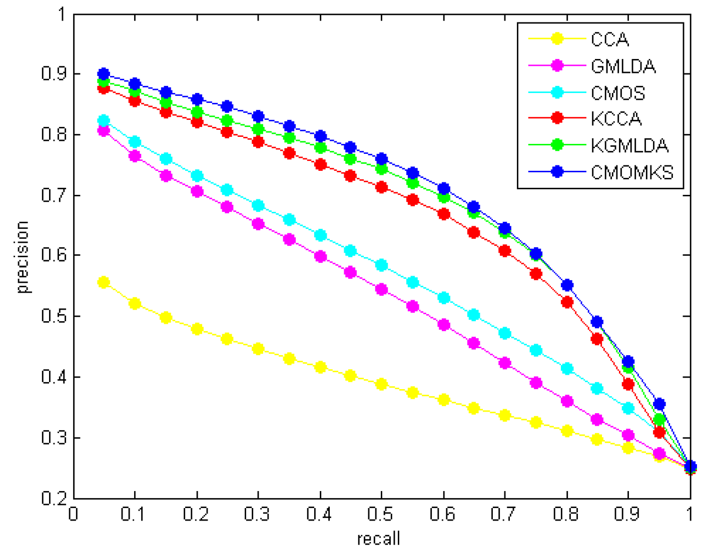
- Datasets: Pascal VOC 2007 consists of annotated consumer photographs collected from Flickr. After removing images without tags, we obtain a training set with 5000 images and a test set with 4919 images.
- Features: Images are represented by ‘conv2’, ‘conv5’, ‘fc6’, ‘fc7’,‘prob’ CNN features. We concatenate CNN features for primal methods, and average kernel matrices for KCCA and KGMLDA. Texts are represented by tag occurrence features.

Performance:

method	img2txt	txt2img	average
CCA	0.346	0.395	0.371
PLS	0.532	0.490	0.511
GMLDA	0.550	0.539	0.545
ml-CCA	0.584	0.572	0.578
LCFS	0.551	0.528	0.540
Bi-CMSRM	0.541	0.516	0.529
SSI	0.576	0.530	0.553
CMOS	0.586	0.600	0.593
KCCA	0.647	0.655	0.651
KGMLDA	0.675	0.676	0.676
CMOMKS	**0.709**	**0.707**	**0.708**



img2txt precision-recall curve



txt2img precision-recall curve

Conclusions:

- We have proposed CMOS and its multiple kernel extension CMOMKS to learn a similarity function between heterogeneous data modalities by preserving relative similarity constraints from two directions.
- The CMOS online model is learned by the Passive-Aggressive algorithm. Multiple kernelized similarity function is further combined in CMOMKS by the Hedging algorithm.
- Experimental results on three public cross-modal datasets have demonstrated that the proposed methods outperform state-of-the-art approaches.