MULTI-VIEW DEEP METRIC LEARNING FOR IMAGE CLASSIFICATION

Dewei Li, Jingjing Tang, Yingjie Tian and Xuchan Ju

{lidewei15, tangjingjing13}@mails.ucas.ac.cn, tyj@ucas.ac.cn

PROBLEM

Image classification is one of the core problems in computer vision. There exist many challenges in the visual contents of images, including intra-class variance, scale and viewpoint variation, background clutter, etc., which bring negative effects to the performance of the current methods.

CONTRIBUTION

A novel framework that combine the techniques of metric learning, multi-view learning and deep learning are proposed to make image classification. Multiple kinds of features are extracted to obtain information from different sides and deep neural networks make nonlinear transformations on these features to gather similar images and scatter dissimilar images.

PRELIMINARIES AND NOTATIONS

Give a multi-view dataset with m training examples from c classes, $T = \{T_v \in R^{n_v \times m} \times Y\}_{v=1}^{V}$, where

$$T_v = \{(x_{v1}, y_1), (x_{v2}, y_2), \cdots, (x_{vm}, y_m)\} \qquad (1)$$

is the feature set from v-th view and $y_i \in Y = \{1, 2, \cdots, c\}$ is the label corresponding to each feature input.

Metric learning: learning a data-dependent metric to measure similarity more precisely

Multi-view learning: incoperate the information from different views

ACKNOWLEDGEMENTS

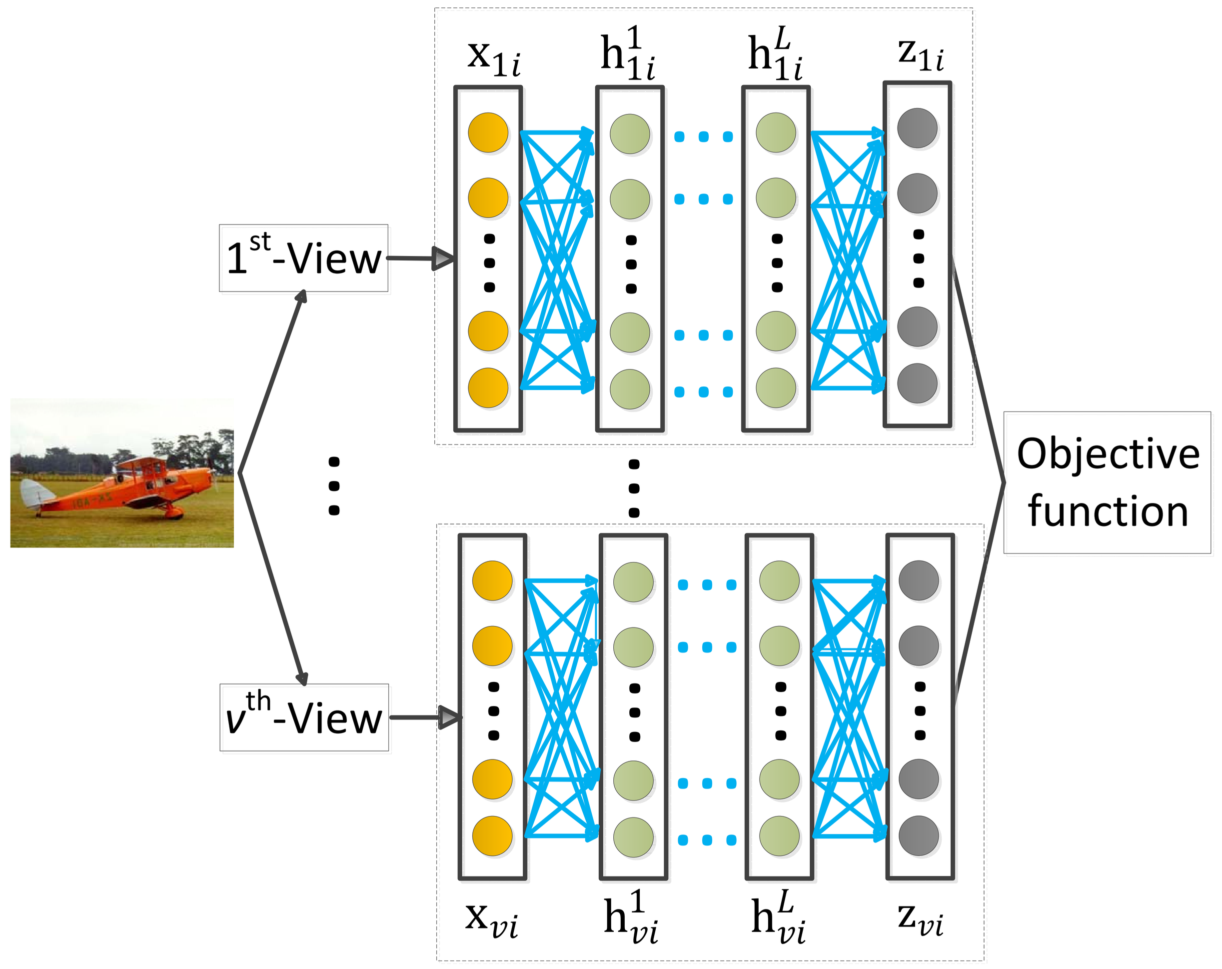
This work has been partially supported by grants from National Natural Science Foundation of China (Nos .61472390, 11271361, 71331005, and 11226089), Major International (Regional) Joint Research Project (No. 71110107026) and the Beijing Natural Science Foundation (No.1162005).

REFERENCES

[1] Lu Jiwen, Wang Gang, Deng Weihong, Moulin Pierre, Zhou Jie Multi-manifold deep metric learning for image set classification In *CVPR'15*, 2015.

[2] Xing, Eric P and Jordan, Michael I and Russell, Stuart and Ng, Andrew Y Distance metric learning with application to clustering with side-information In *NIPS'02*, 2002.

PROPOSED FRAMEWORK



V deep neural networks are constructed, each for a view, to make nonlinear transformation. For each training input x_{vi}, its output of the first layer in the v-th network is $h_{vi}^1 = s(\hat{W}_v^1 \hat{x}_{vi})$, where $\hat{W}_v^1 = (W_v^1, b_v^1), \hat{x}_{vi} = (x_{vi}^\top, 1)^\top$. the output of the top hidden layer is $h_{vi}^L = s(W_v^L h_{vi}^{L-1} + b_v^L) = s(\hat{W}_v^L \hat{h}_{vi}^{L-1})$. Then the output

$$z_{vi} = s(\hat{W}_v^{L+1} \hat{h}_{vi}^L).$$

The output should meet two conditions:
(1) Cohensiveness and scatterness:

$$\min_{\hat{W}} J_1 = \sum_{v=1}^{V} \sum_{i=1}^{m} \alpha_v (d_1(z_{vi}) - C d_2(z_{vi}))$$

where

$$d_1(z_{vi}) = \frac{1}{K} \sum_{z_{vk} \in S_{vi}} \|z_{vi} - z_{vk}\|^2$$

$$d_2(z_{vi}) = \frac{1}{K-1} \sum_{z_{vk} \in D_{vi}} \|z_{vi} - z_{vk}\|^2$$

(2) Consistency:

$$\min J_2 = \sum_{i=1}^{m} \sum_{k,l=1}^{V} d(z_{ki}, z_{li})$$

The framework of multi-view deep metric learning is established by integerating the above two goals:

$$\min J = J_1 + J_2$$

SOLUTION

Alternative optimization is used to obtain the solution alternately. First, the weight α is initialized and fixed, then the object function is an unconstrained problem and gradient descent is adopted to solve problem iteratively. The gradient of the objective function with respect to $\hat{W}_v^l$ is

$$\frac{\partial J}{\partial \hat{W}_v^l} = \alpha_v \sum_{i=1}^{m} \frac{\partial}{\partial \hat{W}_v^l}(d_1 - C d_2) + \frac{\varepsilon}{2} \sum_{i=1}^{m} \sum_{l \neq v} \frac{\partial}{\partial \hat{W}_v^l} d(z_{ki}, z_{li}) + \lambda \hat{W}_v^l \qquad (2)$$

After obtaining the weight matrix $\hat{W}$, then α can be calculated based on the KKT condition,

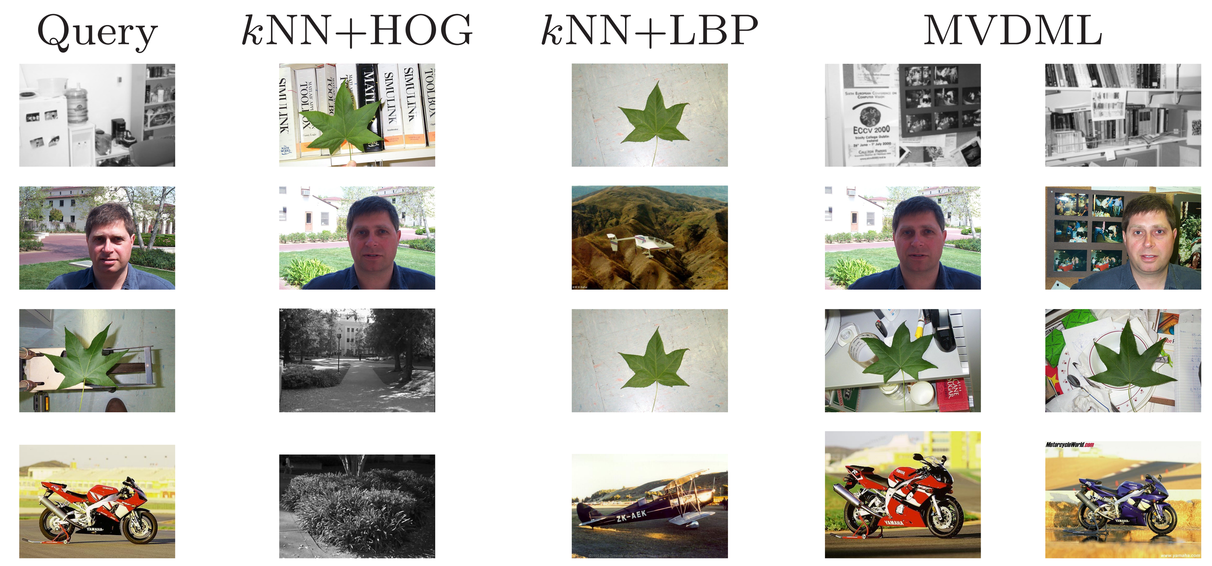
$$\alpha = \frac{\mu e + e^\top \kappa e - V\kappa}{\mu V} \qquad (3)$$

where $\kappa = (\kappa_1, \cdots, \kappa_V) \in R^V$ and $\kappa_v = \sum_{i=1}^{m} (d_1(z_{vi}) - C d_2(z_{vi})), v = 1, \cdots, V$.

PREDICT

Give a test image with V views, all of its views will be input to corresponding networks learned from the training images. Suppose that the outputs are $z_1, z_2, \cdots, z_V$ and their nearest neighbors from the c-th class of the train set can be found, $z_1', z_2', \cdots, z_V'$. The distance between the test image and the nearest neighbor in c-th class is $d^c = \sum_{v=1}^{V} \alpha_v \|z_v - z_v'\|_2^2$. So the label of the test image is $y = \arg\min_c d^c$.

NEAREST NEIGHBORS



CLASSIFICATION AND COMPLEXITY

Datasets	View	Euc	MCML	LMNN	ITML	MVDML
Caltech (600&6)	Single	11.3±5.0	7.7±0.3	8.0±2.2	7.8±2.9	\
	Single	18.2±0.8	15.3±2.8	15.0±2.0	11.8±0.6	\
	Multiple	11.3±5.0	7.0±1.5	7.5±1.7	6.3±1.8	**6.3**±0.3
Galaxy (522&3)	Single	19.2±2.6	14.0±4.4	14.2±3.4	14.0±3.4	\
	Single	20.5±3.4	20.5±3.2	21.1±0.9	15.7±1.3	\
	Multiple	19.2±2.3	13.2±3.0	14.2±2.9	14.0±4.4	**11.7**±0.3
GRAZ02 (800&4)	Single	58.2±3.0	55.3±2.0	53.6±2.7	57.9±2.6	\
	Single	51.3±2.5	50.1±4.1	38.3±2.1	52.5±5.8	\
	Multiple	57.7±3.2	52.5±4.8	48.3±0.6	58.9±3.7	**42.8**±1.5
bike (745&2)	Single	40.1±2.6	30.6±1.9	38.8±3.7	36.8±1.5	\
	Single	31.6±5.4	32.4±1.2	31.3±2.7	31.6±0.8	\
	Multiple	40.1±2.4	**30.2**±2.4	31.7±1.7	35.8±2.0	32.8±2.9
car (800&2)	Single	43.0±4.9	39.5±1.5	42.8±1.8	41.2±4.0	\
	Single	40.4±0.4	39.3±1.1	35.9±3.1	36.7±1.7	\
	Multiple	42.6±4.9	37.7±1.5	**36.5**±1.9	40.7±4.1	38.0±2.1
person (691&2)	Single	36.9±4.5	25.5±1.5	30.3±0.9	31.2±5.8	\
	Single	35.9±3.1	35.2±1.2	32.1±1.1	34.2±3.6	\
	Multiple	36.8±4.5	**27.2**±1.8	28.8±2.0	30.7±5.1	28.4±2.0

Datasets	MCML	LMNN	ITML	MVDML
Caltech	605+685/2056	313+368/533	119+97/154	49s
Galaxy	427+375/1783	93+357/172	119+113/133	44s
GRAZ02	[illegible]	[illegible]	[illegible]	[illegible]