Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Exploring Commonality and Individuality for Multi-Modal Curriculum Learning

Chen Gong[†,⋆]
[†]Pattern Computing and Applications (PCA) Lab, Nanjing University of Science and Technology
[⋆]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University
E-mail: chen.gong@njust.edu.cn

Abstract

Curriculum Learning (CL) mimics the cognitive process of humans and favors a learning algorithm to follow the logical learning sequence from simple examples to more difficult ones. Recent studies show that selecting the simplest curriculum examples from different modalities for graph-based label propagation can yield better performance than simply leveraging single modality. However, they forcibly require the curriculums generated by all modalities to be identical to a common curriculum, which discard the individuality of every modality and produce the inaccurate curriculum for the subsequent learning. Therefore, this paper proposes a novel multi-modal CL algorithm by comprehensively investigating both the individuality and commonality of different modalities. By considering the curriculums of multiple modalities altogether, their common preference on selecting the simplest examples can be explored by a row-sparse matrix, and their distinct opinions are captured by a sparse noise matrix. As a consequence, a "soft" fusion of multiple curriculums from different modalities is achieved and the propagation quality can thus be improved. Comprehensive empirical studies reveal that our method can generate higher accuracy than the state-of-the-art multi-modal CL approach and label propagation algorithms on various image classification tasks.

Introduction

Curriculum Learning (CL) (Bengio et al. 2009) advocates logically training a classifier by gradually leveraging the examples from simple to difficult. In contrast to massive existing classifiers (*e.g.* Support Vector Machines and Naive Bayesian Classifier) that are trained on all examples at one time, CL establishes a sequence of curriculums so that only the optimal curriculum containing the simplest examples are invoked to train the classifier in each learning round. Such "starting small" strategy is very similar to the human's knowledge acquisition process from childhood to adulthood, and also has been demonstrated to be effective in machine learning (Kumar, Packer, and Koller 2010; Jiang et al. 2015; Gong et al. 2016a) and computer vision (Lee and Grauman 2011; Supancic and Ramanan 2013; Gong et al. 2015; Pentina, Sharmanska, and Lampert 2015).

The concept of curriculum learning was originally proposed by Bengio et al. (Bengio et al. 2009). After that, CL was usually realized under two frameworks: Self-Paced Learning (SPL) and Teaching-to-Learn and Learning-to-Teach (TLLT). SPL was formally developed in (Kumar, Packer, and Koller 2010), which employs latent SVMs as a learner and considers an example as simple if it lies far from the margin. Apart from selecting the simple examples for training, Jiang et al. (Jiang et al. 2014b) also require that the selected curriculum examples to be diverse. Furthermore, they also favor of harnessing the dynamic example difficulty revealed during learning in addition to the estimation of difficulty before learning (Jiang et al. 2015). Up to now, SPL has been widely used in various learning problems such as clustering (Xu, Tao, and Xu 2015), domain adaptation (Tang et al. 2012a), dictionary learning (Tang et al. 2012b), and zero-shot learning (Jiang et al. 2014a).

The TLLT framework (Gong et al. 2016a) was specifically designed for graph-based label propagation (Zhu and Ghahramani 2002). It is composed of two stages named teaching-to-learn and learning-to-teach. In teaching-to-learn, the "teacher" (*i.e.* a teaching algorithm) chooses the simplest examples for the "learner" (*i.e.* a propagation algorithm) by assessing their reliability and discriminability. In learning-to-teach, the learner delivers a learning feedback to the teacher to help it decide the suitable curriculum for the next learning round. This basic TLLT framework has been further extended to multi-label cases (Gong et al. 2016c) and multi-modal cases (Gong et al. 2016b).

The multi-modal setting assumes that every example can be characterized by different modalities (Xu, Tao, and Xu 2013). As the pioneering work of adapting CL to multi-modal setting, Gong et al. (Gong et al. 2016b) demonstrate that integrating the curriculums generated by different modalities helps to improve the classification accuracy. This is because the curriculums from various modalities can complement to each other to yield an overall good curriculum. However, this method directly minimizes the error between the curriculum of every modality and the central optimal curriculum, which is a "hard" constraint suppressing the individuality possessed by every modality. Such imperfect fusion scheme degrades the curriculum quality and is unfavorable to obtaining satisfactory classification results. To address this defect, this paper explicitly models the commonality among all modalities as well as their individualities to achieve a "soft" curriculum fusion, so our algo-

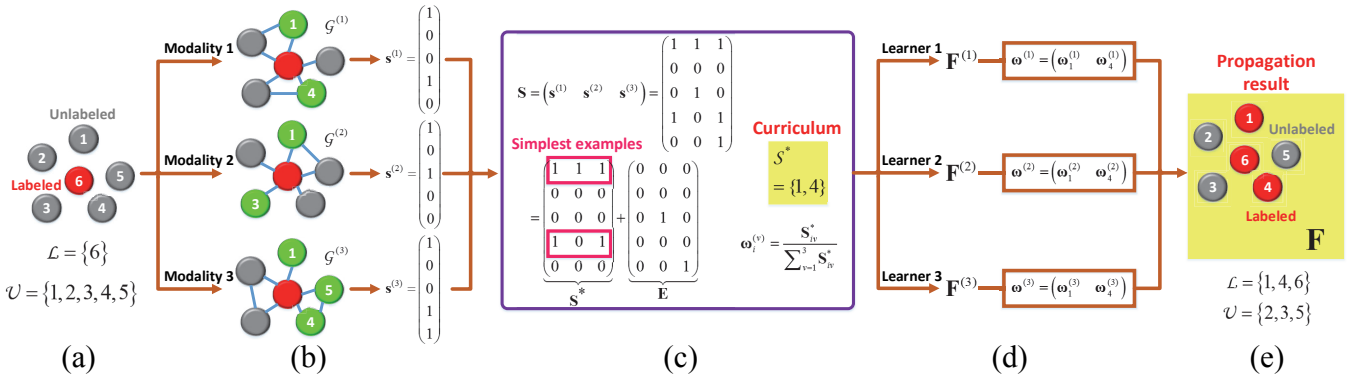

Figure 1: The framework of our SMMCL algorithm. In (a), the examples in labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$ are denoted by red and gray balls, respectively. In (b), the v-th ($v = 1, 2, 3$ in this figure) teachers should choose the simplest examples (green balls) from their own modalities based on the graphs $\mathcal{G}^{(v)}$, and the selected curriculum examples are encoded in the selection vectors $\mathbf{s}^{(v)}$ whose length equals to the size of $\mathcal{U}$. In (c), the selection vectors produced by all teachers are put together to form a stacked matrix $\mathbf{S}$, which can be regarded as the sum of a row-sparse matrix $\mathbf{S}^*$ and a noise matrix $\mathbf{E}$. The non-zero rows of $\mathbf{S}^*$ indicated by the magenta boxes correspond to the simplest examples that should be taken into the curriculum $\mathcal{S}^*$. Besides, the (i, v)-th element of $\mathbf{S}^*$ indicates the weight $\boldsymbol{\omega}_i^{(v)}$ of the v-th modality on deciding the label of the i-th curriculum example. In (d), the learners (*i.e.* propagation algorithms) "learn" the examples in $\mathcal{S}^*$ by propagating the label information to them, and the resultant label matrices are $\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \mathbf{F}^{(3)}$. In (e), a unified label matrix $\mathbf{F}$ is obtained by adding $\mathbf{F}^{(1)} \sim \mathbf{F}^{(3)}$ weighted by $\boldsymbol{\omega}^{(1)} \sim \boldsymbol{\omega}^{(3)}$. The labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$ are also updated accordingly.

rithm is termed "Soft Multi-Modal Curriculum Learning" (SMMCL). Specifically, we assume that the curriculums of multiple modalities as a whole can be decomposed as a row-sparse component plus a noise component, in which the row-sparse component describes the commonality shared by multiple modalities and the noise component captures the individuality carried out by each of the modalities. As a result, the involved modalities are more easily to reach an agreement on selecting the simplest examples, and the selected curriculum examples are also more accurate than those produced by (Gong et al. 2016b).

Framework of Our Method

This section briefly introduces the framework of the proposed SMMCL algorithm. Given totally $n = l + u$ examples $\mathcal{X} = \{\mathbf{x}_1 \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$ where the first l examples constitute the labeled set $\mathcal{L}$ and the last u examples form the unlabeled set $\mathcal{U}$ (see Fig. 1(a)), the task of graph-based label propagation is to iteratively propagate the known labels $\{y_i\}_{i=1}^{l}$ of $\mathcal{L}$ to $\mathcal{U}$.

For multi-modal cases, we assume that each example $\mathbf{x}_i$ can be characterized by V different modalities, so V graphs $\mathcal{G}^{(1)}, \cdots, \mathcal{G}^{(V)}$ can be built correspondingly (see Fig. 1(b)). In these graphs, the vertices represent n examples and the edges depict the similarities between these examples. Similar to (Gong et al. 2016b), we also associate each modality with a teacher and a learner, and in each learning round the V teachers should pick up the overall simplest examples (denoted by the set $\mathcal{S}^*$) for the stepwise learners. To this end, the v-th ($v = 1, \cdots, V$) teacher should generate an optimal curriculum from its own viewpoint that is recorded by a $\{0, 1\}$-binary selection vector $\mathbf{s}^{(v)}$. Here $\mathbf{s}_i^{(v)} = 1$ if the i-th example is considered simple and is chosen by the v-th teacher, and $\mathbf{s}_i^{(v)} = 0$ otherwise. After that, the decisions made by all V teachers are integrated into a unified curriculum $\mathcal{S}^*$ (see Fig. 1(c)), during which the commonality of the teachers and their individualities are discovered by the row-sparse matrix $\mathbf{S}^*$ and sparse noise matrix $\mathbf{E}$ accordingly. Given $\mathcal{S}^*$, the V learners will classify the examples in $\mathcal{S}^*$ by respectively propagating the labels from $\mathcal{L}$ to $\mathcal{S}^*$ from V modalities, and the obtained results are recorded in the label matrices $\mathbf{F}^{(v)} \in \mathbb{R}^{n \times c}$ ($v = 1, \cdots, V$, and c is the number of classes) (see Fig. 1(d)). The i-th row of $\mathbf{F}^{(v)}$ is the label vector of the example $\mathbf{x}_i$ with its j-th element (*i.e.* the (i, j)-th element of $\mathbf{F}^{(v)}$) encoding the probability of the i-th example belonging to the j-th ($j = 1, \cdots, c$) class. Finally, the propagation results $\mathbf{F}^{(1)}, \cdots, \mathbf{F}^{(V)}$ are fused into $\mathbf{F}$ by considering their weights $\boldsymbol{\omega}^{(v)}$ on all the curriculum examples (see Fig. 1(e)). The labeled set and unlabeled set are then updated by $\mathcal{L} := \mathcal{L} \cup \mathcal{S}^*$ and $\mathcal{U} := \mathcal{U} - \mathcal{S}^*$, respectively. Such teaching and learning process iterates until the set $\mathcal{U} = \varnothing$.

The most critical step of our SMMCL algorithm is how to make the teachers maximally reach an agreement on determining suitable $\mathcal{S}^*$ based on their individual decisions (*i.e.* Fig. 1(c)). Therefore, we propose a novel multi-modal teaching algorithm that will be detailed in the next section.

Model Description

According to (Gong et al. 2016b), the difficulty level of an example $\mathbf{x}_i \in \mathcal{U}$ under single modality can be evaluated by its reliability and discriminability. By taking y_i as a random variable of $\mathbf{x}_i$ and treating the propagations on $\mathcal{G}^{(v)}$ as a Gaussian process over the random vector $\mathbf{y} =$

$(y_1, \cdots, y_n)^\top$, the reliability is modeled by the conditional entropy $H(y_i|\mathbf{y}_\mathcal{L})$, where $\mathbf{y}_\mathcal{L}$ is the subvector of $\mathbf{y}$ corresponding to the labeled set $\mathcal{L}$. Therefore, similar to the derivations in (Gong et al. 2016b), we have

$$H(y_i|\mathbf{y}_\mathcal{L}) \propto \left|\mathbf{\Sigma}_{i|\mathcal{L}}\right| = \mathbf{\Sigma}_{i,i} - \mathbf{\Sigma}_{i,\mathcal{L}}\mathbf{\Sigma}_{\mathcal{L},\mathcal{L}}^{-1}\mathbf{\Sigma}_{\mathcal{L},i}, \quad (1)$$

where $\mathbf{\Sigma}$ is the covariance matrix of the random vector $\mathbf{y}$, and $\mathbf{\Sigma}_{i,i}, \mathbf{\Sigma}_{i,\mathcal{L}}, \mathbf{\Sigma}_{\mathcal{L},i}, \mathbf{\Sigma}_{\mathcal{L},\mathcal{L}}$ are submatrices of $\mathbf{\Sigma}$ associated with the corresponding subscripts. The covariance matrix $\mathbf{\Sigma}$ is defined by $\mathbf{\Sigma} = (\mathbf{L}+\mathbf{I}/\kappa^2)^{-1}$ where $\mathbf{L}$ is graph Laplacian (Zhu, Ghahramani, and Lafferty 2003), $\mathbf{I}$ is identity matrix, and κ^2 is the parameter fixed to 100 throughout this paper. Small $H(y_i|\mathbf{y}_\mathcal{L})$ means that classifying $\mathbf{x}_i$ is reliable and it should be incorporated by the curriculum $\mathcal{S}^*$.

The discriminability of $\mathbf{x}_i$ depicts its tendency belonging to a certain class, which is modeled by the difference of average commute time (Qiu and Hancock 2007) from $\mathbf{x}_i$ to its two nearest classes $\mathcal{C}_1$ and $\mathcal{C}_2$, namely

$$M(\mathbf{x}_i) = \bar{T}(\mathbf{x}_i, \mathcal{C}_2) - \bar{T}(\mathbf{x}_i, \mathcal{C}_1), \quad (2)$$

where $\bar{T}(\mathbf{x}_i, \mathcal{C}_j)$ computes the average commute time between $\mathbf{x}_i$ and all the examples of class $\mathcal{C}_j$ $(j = 1, 2)$. Large $M(\mathbf{x}_i)$ means that $\mathbf{x}_i$ is significantly inclined to the class $\mathcal{C}_1$ and thus it is ideal to be a curriculum example.

By taking account of $\mathbf{x}_i$'s reliability and discriminability together, the difficulty of $\mathbf{x}_i$ in terms of the v-th modality (*i.e.* $R_i^{(v)}$) is then represented by

$$R_i^{(v)} = H(y_i|\mathbf{y}_\mathcal{L}) + 1/M(\mathbf{x}_i). \quad (3)$$

The example with small $R_i^{(v)}$ is simple and is suitable for the current propagation conducted by the v-th learner.

However, different teachers often have different selections on the simplest examples, as the difficulties of an example revealed by different modalities are distinct. Therefore, we should find the common curriculum examples that are maximally agreed by all V teachers. To this end, we put the binary selection vectors $\mathbf{s}^{(v)}$ $(v = 1, \cdots, V)$ of the V teachers altogether as a matrix $\mathbf{S} = \left(\mathbf{s}^{(1)}, \cdots, \mathbf{s}^{(V)}\right)$, then its all-zero rows will indicate the difficult examples considered by all V teachers. Practically, the teachers can hardly draw the identical conclusion on deciding $\mathcal{S}^*$, so we assume that $\mathbf{S}$ implicitly contains a row-sparse component $\mathbf{S}^*$ representing the consensus of all teachers, and a sparse noise term $\mathbf{E}$ capturing the individuality of each modality. As a result, we have $\mathbf{S} = \mathbf{S}^* + \mathbf{E}$ with $\mathbf{S}$, $\mathbf{S}^*$ and $\mathbf{E}$ being $\{0, 1\}$-binary matrices (see Fig. 1(c)). Thereby, our model is formulated as

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{S}^*,\mathbf{E}} \sum\nolimits_{v=1}^{V} \mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} + \alpha\|\mathbf{S}^*\|_{2,1} + \beta\|\mathbf{E}\|_1 \\ &s.t.\ \mathbf{S} = \mathbf{S}^* + \mathbf{E},\ \mathbf{S}_{ij} \in \{0,1\},\ \mathbf{S}^*_{ij} \in \{0,1\},\ \mathbf{E}_{ij} \in \{0,1\}, \\ &\quad \mathbf{1}^\top\mathbf{s}^{(v)} = Q,\ \forall\, v = 1, 2, \cdots, V \end{aligned} \quad (4)$$

where $\mathbf{R}^{(v)}$ is a diagonal matrix with the i-th diagonal element being $R_i^{(v)}$ defined in Eq. (3), $\|\mathbf{S}^*\|_{2,1} = \sum_i \sqrt{\sum_j \mathbf{S}^{*2}_{ij}}$ computes $\mathbf{S}^*$'s $l_{2,1}$ norm (Chang and Yang 2016; Chang et al. 2015), $\|\mathbf{E}\|_1 = \sum_{i,j} |\mathbf{E}_{ij}|$ is the l_1 norm of matrix $\mathbf{E}$, $\mathbf{1}$ is the all-one column vector with the same length as $\mathbf{s}^{(v)}$, and $\alpha, \beta, Q > 0$ are free parameters.

In the objective function of Eq. (4), the first term governs the example selection of every single modality. Minimizing this term requires $\mathbf{s}_i^{(v)}$ to be small if the difficulty value $R_i^{(v)}$ of $\mathbf{x}_i$ large, which suggests that $\mathbf{x}_i$ should not be a curriculum example. By utilizing the $l_{2,1}$ norm of $\mathbf{S}^*$, the second term exploits the common decision made by all the teachers from different modalities. The third term models the unique opinion of each teacher and minimizing it drives all teachers to reach an agreement as possible as they can. The constraints $\mathbf{1}^\top\mathbf{s}^{(v)} = Q$ $(v = 1, \cdots, V)$ ensure that the simplest examples recommended by V teachers are not skewed.

However, Eq. (4) is difficult to solve because of the binary constraints, so we relax these constraints by modifying the objective function, and obtain

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{S}^*,\mathbf{E}} \sum\nolimits_{v=1}^{V} \mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} + \alpha\|\mathbf{S}^*\|_{2,1} + \beta\|\mathbf{E}\|_1 \\ &+\frac{\gamma}{2}\left(\|\mathbf{S}\circ\mathbf{S}-\mathbf{S}\|_{\mathrm{F}}^2+\|\mathbf{S}^*\circ\mathbf{S}^*-\mathbf{S}^*\|_{\mathrm{F}}^2+\|\mathbf{E}^*\circ\mathbf{E}^*-\mathbf{E}^*\|_{\mathrm{F}}^2\right), \\ &s.t.\ \mathbf{S} = \mathbf{S}^* + \mathbf{E},\quad \mathbf{1}^\top\mathbf{s}^{(v)} = Q,\ \forall\, v = 1, 2, \cdots, V \end{aligned} \quad (5)$$

where "$\|\cdot\|_{\mathrm{F}}$" denotes the Frobenius norm, "$\circ$" represents the Hadamard product that is $(\mathbf{A}\circ\mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$, and γ is the weighting parameter. When Eq. (5) is solved, the non-zero rows of $\mathbf{S}^*$ will indicate the selected curriculum examples that should be classified by the learners.

Compared with the model in (Gong et al. 2016b) that rashly forces all $\mathbf{s}^{(v)}$ $(v = 1, \cdots, V)$ to a compromised $\mathbf{s}^*$ by minimizing $\left\|\mathbf{s}^{(v)} - \mathbf{s}^*\right\|_2^2$, Eq. (5) developed here tries to discover the underlying consensus among different teachers as well as explicitly preserves the individuality of every teacher, so it achieves "soft" fusion of multiple modalities without loosing their specialities.

Optimization

The problem (5) can be solved via the Alternating Direction Method of Multipliers (ADMM), which alternatively optimizes one variable at one time with the other variables remaining fixed. To decouple the variables $\mathbf{S}^*$ and $\mathbf{S}$, we introduce an auxiliary variable $\mathbf{J}$ and a related constraint $\mathbf{J} = \mathbf{S}^*$, and the original optimization problem (5) is reformulated as

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{S}^*,\mathbf{E},\mathbf{J}} \sum\nolimits_{v=1}^{V} \mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} + \alpha\|\mathbf{J}\|_{2,1} + \beta\|\mathbf{E}\|_1 \\ &+\frac{\gamma}{2}\left(\|\mathbf{S}\circ\mathbf{S}-\mathbf{S}\|_{\mathrm{F}}^2+\|\mathbf{S}^*\circ\mathbf{S}^*-\mathbf{S}^*\|_{\mathrm{F}}^2+\|\mathbf{E}^*\circ\mathbf{E}^*-\mathbf{E}^*\|_{\mathrm{F}}^2\right)\cdot \\ &s.t.\ \mathbf{S} = \mathbf{S}^* + \mathbf{E},\ \mathbf{J} = \mathbf{S}^*,\ \mathbf{1}^\top\mathbf{s}^{(v)} = Q,\ \forall\, v = 1, 2, \cdots, V \end{aligned} \quad (6)$$

Therefore, the augmented Lagrangian function is

$$\begin{aligned}&\sum_{v=1}^{V}\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)}+\alpha\|\mathbf{J}\|_{2,1}+\beta\|\mathbf{E}\|_1\\&+\frac{\gamma}{2}\big(\|\mathbf{S}\circ\mathbf{S}-\mathbf{S}\|_{\mathrm{F}}^2+\|\mathbf{S}^*\circ\mathbf{S}^*-\mathbf{S}^*\|_{\mathrm{F}}^2+\|\mathbf{E}^*\circ\mathbf{E}^*-\mathbf{E}^*\|_{\mathrm{F}}^2\big)\\&+tr\big(\mathbf{\Lambda}_1^\top(\mathbf{S}-\mathbf{S}^*-\mathbf{E})\big)+tr\big(\mathbf{\Lambda}_2^\top(\mathbf{J}-\mathbf{S}^*)\big)+\sum_{v=1}^{V}\tau_v(\mathbf{1}^\top\mathbf{s}^{(v)}-Q)\\&+\frac{\mu}{2}\Big[\|\mathbf{S}-\mathbf{S}^*-\mathbf{E}\|_{\mathrm{F}}^2+\|\mathbf{J}-\mathbf{S}^*\|_{\mathrm{F}}^2+\sum_{v=1}^{V}(\mathbf{1}^\top\mathbf{s}^{(v)}-Q)^2\Big]\end{aligned},\tag{7}$$

where $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \tau_v$ are Lagrangian multipliers, and $\mu > 0$ is the penalty coefficient. Based on Eq. (7), the variables $\mathbf{S}, \mathbf{S}^*, \mathbf{E}, \mathbf{J}$ can be sequentially updated via an iterative way.

Update J: The subproblem related to $\mathbf{J}$ is

$$\min_{\mathbf{J}}\ \alpha\|\mathbf{J}\|_{2,1}+tr\big(\mathbf{\Lambda}_2^\top(\mathbf{J}-\mathbf{S}^*)\big)+\frac{\mu}{2}\|\mathbf{J}-\mathbf{S}^*\|_{\mathrm{F}}^2,\tag{8}$$

which is equivalent to

$$\min_{\mathbf{J}}\ \frac{\alpha}{\mu}\|\mathbf{J}\|_{2,1}+\frac{1}{2}\left\|\mathbf{J}-\big(\mathbf{S}^*-\frac{1}{\mu}\mathbf{\Lambda}_2\big)\right\|_{\mathrm{F}}^2,\tag{9}$$

of which the optimal solution is (Liu et al. 2012)

$$\mathbf{J}_{i,:}=\begin{cases}\dfrac{\|\mathbf{T}_{i,:}\|_2-\alpha/\mu}{\|\mathbf{T}_{i,:}\|_2}\mathbf{T}_{i,:}, & \alpha/\mu<\|\mathbf{T}_{i,:}\|_2\\ 0, & \text{otherwise}\end{cases},\tag{10}$$

where $\mathbf{T}=\mathbf{S}^*-\frac{1}{\mu}\mathbf{\Lambda}_2$ and $\mathbf{T}_{i,:}$ denotes the i-th row of $\mathbf{T}$.

Update $\mathbf{S}^*$: By denoting $\mathbf{M}_1=\mathbf{S}^*\circ\mathbf{S}^*$, the subproblem regarding $\mathbf{S}^*$ is

$$\begin{aligned}\min_{\mathbf{S}^*}\ &\frac{\gamma}{2}\|\mathbf{S}^*-\mathbf{M}_1\|_{\mathrm{F}}^2+tr\big(\mathbf{\Lambda}_1^\top(\mathbf{S}-\mathbf{S}^*-\mathbf{E})\big)+tr\big(\mathbf{\Lambda}_2^\top(\mathbf{J}-\mathbf{S}^*)\big)\\&+\frac{\mu}{2}\left[\|\mathbf{S}-\mathbf{S}^*-\mathbf{E}\|_{\mathrm{F}}^2+\|\mathbf{J}-\mathbf{S}^*\|_{\mathrm{F}}^2\right]\end{aligned}.\tag{11}$$

By calculating the derivative of above objective to $\mathbf{S}^*$, and then setting the result to 0, the closed-form solution for $\mathbf{S}^*$ is obtained by

$$\mathbf{S}^*=\frac{1}{\gamma+2\mu}\left[\mathbf{\Lambda}_1+\mathbf{\Lambda}_2-\mu(\mathbf{E}-\mathbf{S}-\mathbf{J})+\gamma\mathbf{M}_1\right].\tag{12}$$

Update E: By denoting $\mathbf{M}_2=\mathbf{E}\circ\mathbf{E}$, the subproblem for optimizing $\mathbf{E}$ is

$$\begin{aligned}\min_{\mathbf{E}}\ &\beta\|\mathbf{E}\|_1+\frac{\gamma}{2}\|\mathbf{E}-\mathbf{M}_2\|_{\mathrm{F}}^2\\&+tr\big(\mathbf{\Lambda}_1^\top(\mathbf{S}-\mathbf{S}^*-\mathbf{E})\big)+\frac{\mu}{2}\|\mathbf{S}-\mathbf{S}^*-\mathbf{E}\|_{\mathrm{F}}^2\end{aligned}.\tag{13}$$

After ignoring the constant variables and re-arranging Eq. (13), the $\mathbf{E}$-subproblem is formed as

$$\min_{\mathbf{E}}\ \beta\|\mathbf{E}\|_1+\frac{\gamma+\mu}{2}\left\|\mathbf{E}-\frac{1}{\gamma+\mu}\mathbf{B}\right\|_{\mathrm{F}}^2,\tag{14}$$

Algorithm 1 The ADMM process for solving Eq. (5)

Input: $\mathbf{R}^{(v)}$, α, β, γ, Q, $\mu=1$, $\mu_{max}=10^8$, $\rho=1.2$, $\epsilon=10^{-4}$, $MaxIter=50$, initial $\mathbf{S}$, $\mathbf{E}$, $\mathbf{S}^*$.
$iter=0$;
repeat
  $\mathbf{M}_1=\mathbf{S}^*\circ\mathbf{S}^*$, $\mathbf{M}_2=\mathbf{E}\circ\mathbf{E}$, $\mathbf{M}_3=\mathbf{S}\circ\mathbf{S}$;
  Update $\mathbf{J}$ via Eq. (10);
  Update $\mathbf{S}^*$ via Eq. (12);
  Update $\mathbf{E}$ via Eq. (15);
  for $v=1$ to V **do**
    Update $\mathbf{s}^{(v)}$ via Eq. (17);
  end for
  $\mathbf{\Lambda}_1:=\mathbf{\Lambda}_1+\mu(\mathbf{S}-\mathbf{S}^*-\mathbf{E})$, $\mathbf{\Lambda}_2:=\mathbf{\Lambda}_2+\mu(\mathbf{J}-\mathbf{S}^*)$, $\tau_v:=\tau_v+\mu(\mathbf{1}^\top\mathbf{s}^{(v)}-Q),\forall\, v=1,\cdots,V$;
  $\mu=\min(\rho\mu,\mu_{max})$;
  $iter:=iter+1$;
until $\left\|\mathbf{S}^{*(iter)}-\mathbf{S}^{*(iter-1)}\right\|_{\mathrm{F}}^2\le\epsilon$ or $iter=MaxIter$
Output: The optimal $\mathbf{S}$, $\mathbf{S}^*$, $\mathbf{E}$.

where $\mathbf{B}=\frac{1}{\gamma+\mu}[\gamma\mathbf{M}_2+\mathbf{\Lambda}_1-\mu(\mathbf{S}^*-\mathbf{S})]$. By employing the soft-thresholding operator (Lin, Chen, and Ma 2010), the solution of Eq. (14) is expressed as

$$\mathbf{E}_{ij}=\begin{cases}\mathbf{B}_{ij}-\dfrac{\beta}{\gamma+\mu}, & \mathbf{B}_{ij}>\dfrac{\beta}{\gamma+\mu}\\ \mathbf{B}_{ij}+\dfrac{\beta}{\gamma+\mu}, & \mathbf{B}_{ij}<\dfrac{-\beta}{\gamma+\mu}\\ 0, & \text{otherwise}\end{cases}.\tag{15}$$

Update $\mathbf{s}^{(v)}$: Note that the columns of $\mathbf{S}$ (*i.e.* $\mathbf{s}^{(v)}$) are independent to each other in our problem (6), so they can be updated separately. Suppose $\mathbf{M}_3=\mathbf{S}\circ\mathbf{S}$, then the objective function regarding $\mathbf{s}^{(v)}$ is

$$\begin{aligned}\min_{\mathbf{s}^{(v)}}\ &\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)}+\frac{\gamma}{2}\left\|\mathbf{s}^{(v)}-\mathbf{M}_3^v\right\|_2^2+\tau_v(\mathbf{1}^\top\mathbf{s}^{(v)}-Q)\\&+\mathbf{\Lambda}_1^{v\top}\mathbf{s}^{(v)}+\frac{\mu}{2}\left[\left\|\mathbf{s}^{(v)}-\mathbf{S}^{*v}-\mathbf{E}^v\right\|_2^2+(\mathbf{1}^\top\mathbf{s}^{(v)}-Q)^2\right]\end{aligned},\tag{16}$$

where $\mathbf{M}_3^v, \mathbf{\Lambda}_1^v, \mathbf{S}^{*v}, \mathbf{E}^v$ with superscript "v" denote the v-th column of the corresponding matrix, and $\mathbf{R}^{(v)}$ is a diagonal matrix that has appeared in Eq. (4). The solution of Eq. (16) can be easily obtained by setting the derivative of Eq. (16) to $\mathbf{s}^{(v)}$ to 0, which leads to

$$\begin{aligned}\mathbf{s}^{(v)}=&\left[2\mathbf{R}^{(v)}+(\gamma+\mu)\mathbf{I}+\mu\mathbf{1}\mathbf{1}^\top\right]^{-1}\\&\left[\gamma\mathbf{M}_3^v-\mathbf{\Lambda}_1^v+(\mu Q-\tau_v)\mathbf{1}+\mu(\mathbf{S}^{*v}+\mathbf{E}^v)\right].\end{aligned}\tag{17}$$

The entire iterative process for solving Eq. (5) is summarized in Algorithm 1. Its convergence has been theoretically proved in (Lin, Chen, and Ma 2010; Chang et al. 2014) and will be empirically illustrated by the experiments.

Label Propagation and Label Fusion

Given totally s curriculum examples $\mathcal{S}^*=\{\mathbf{x}_1^*,\mathbf{x}_2^*,\cdots,\mathbf{x}_s^*\}$ decided by the teachers, the existing

propagation algorithm Gaussian Field and Harmonic Functions (GFHF) (Zhu and Ghahramani 2002) is employed as learners that propagate the labels from $\mathcal{L}$ to $\mathcal{S}^*$ under different modalities. Then the V label matrices $\mathbf{F}^{(1)}, \cdots, \mathbf{F}^{(V)}$ are combined into a consistent $\mathbf{F}$ as shown in Figs. 1(d)(e). For the t-th propagation, the iterative expression for a specific modality v is:

$$\mathbf{F}_{i,:}^{(v)[t]} = \begin{cases} \mathbf{P}_{i,:}^{(v)} \mathbf{F}^{[t-1]}, & \mathbf{x}_i \in (\mathcal{S}^{*[1:t-1]}) \cup \mathcal{S}^{*[t]} \\ \mathbf{F}_{i,:}^{[0]}, & \mathbf{x}_i \in \mathcal{L}^{[0]} \cup (\mathcal{U}^{[0]} - \mathcal{S}^{*[1:t]}) \end{cases} \tag{18}$$

where $\mathbf{F}_{i,:}^{(v)[t]}$ denotes the i-th row of the matrix $\mathbf{F}^{(v)[t]}$, $\mathbf{F}^{[t-1]}$ is the consistent label matrix produced by the previous propagation, $\mathbf{P}_{i,:}^{(v)}$ represents the i-th row of the *transition matrix* $\mathbf{P}^{(v)}$ calculated by $\mathbf{P}^{(v)} = \mathbf{D}^{(v)-1}\mathbf{W}^{(v)}$. Here $\mathbf{W}^{(v)}$ is the adjacency matrix of graph $\mathcal{G}^{(v)}$ and $\mathbf{D}^{(v)}$ is the corresponding diagonal degree matrix with the diagonal elements defined by $\mathbf{D}_{ii}^{(v)} = \sum_{j=1}^{n} \mathbf{W}_{ij}^{(v)}$. $\mathcal{S}^{*[1:t]} = \mathcal{S}^{*[1]} \cup \cdots \cup \mathcal{S}^{*[t]}$ and $\mathcal{U}^{[0]} - \mathcal{S}^{*[1:t]}$ is the complementary set of $\mathcal{S}^{*[1:t]}$ in $\mathcal{U}^{[0]}$. The superscript "$[t]$" represents the t-th propagation. Such propagation strategy is suggested by Zhu *et al.* (Zhu and Ghahramani 2002) and is identical to the propagation model in (Gong et al. 2016b). The initial state for $\mathbf{x}_i$'s label vector $\mathbf{F}_{i,:}^{[0]}$ is

$$\mathbf{F}_{i,:}^{[0]} = \begin{cases} \underbrace{(1/c, \cdots, 1/c)}_{c}, & \mathbf{x}_i \in \mathcal{U}^{[0]} \\ \left(0, \cdots, \underset{\substack{\downarrow \\ j-th\ element}}{1}, \cdots, 0\right), & \mathbf{x}_i \in \mathcal{C}_j \in \mathcal{L}^{[0]} \end{cases} \tag{19}$$

where c is the total number of classes.

To fuse the label matrices $\mathbf{F}^{(1)}, \cdots, \mathbf{F}^{(V)}$ into a unified $\mathbf{F}$, we should find the weights of these V modalities on deciding the curriculum examples. Specifically, we relate $\boldsymbol{\omega}_i^{(v)}$, which is the weight of $\mathbf{F}_{i,:}^{(v)}$ on $\mathbf{x}_i$, to the tendency of the v-th teacher to choose $\mathbf{x}_i$ as a curriculum example, and consider that the examples strongly recommended by the v-th teacher can be reliably "learned" by the v-th learner. This is because the strong recommendation from the v-th teacher indicates that these examples are quite simple for the v-th learner, therefore the learning result $\mathbf{F}^{(v)}$ is trustable and should be emphasized. Fortunately, the i-th element of the selection vector $\mathbf{s}^{(v)}$ (*i.e.* the (i, v)-th element of matrix $\mathbf{S}^*$) exactly reflects the recommendation level of the v-th teacher on the example $\mathbf{x}_i$. Therefore, the weight $\boldsymbol{\omega}_i^{(v)}$ can be computed by

$$\boldsymbol{\omega}_i^{(v)} = \frac{\mathbf{S}_{iv}^*}{\sum_{v=1}^{V} \mathbf{S}_{iv}^*}, \tag{20}$$

based on which the integrated label vector of the i-th example is derived as

$$\mathbf{F}_{i,:} = \sum\nolimits_{v=1}^{V} \boldsymbol{\omega}_i^{(v)} \mathbf{F}_{i,:}^{(v)}, \tag{21}$$

where the superscript "$[t]$" has been dropped for simplicity.

Algorithm 2 SMMCL for graph-based label propagation

Input: l labeled examples $\mathcal{L} = \{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ with known labels $y_1, \cdots, y_l$ expressed in V modalities; u unlabeled examples $\mathcal{U} = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$ with unknown labels $y_{l+1}, \cdots, y_{l+u}$; Parameters α, β, γ, Q, θ;
// Pre-processing
Construct graphs $\mathcal{G}^{(v)}$ $(v = 1, \cdots, V)$ via (Karasuyama and Mamitsuka 2013a); Compute $\mathbf{R}^{(v)}$ via Eq. (3);
// Multi-modal curriculum generation and propagation
repeat
  Establish the optimal curriculum $\mathcal{S}^*$ by solving Eq. (5) (Algorithm 1);
  Compute the label matrix $\mathbf{F}^{(v)}$ via Eq. (18);
  Compute the weights $\boldsymbol{\omega}_i^{(v)}$ $(v = 1, \cdots, V,\ i = 1, \cdots, s)$ via Eq. (20);
  Fuse V label matrices to $\mathbf{F}$ via Eq. (21);
  $\mathcal{L} := \mathcal{L} \cup \mathcal{S}^*$; $\mathcal{U} := \mathcal{U} - \mathcal{S}^*$;
until $\mathcal{U} = \varnothing$;
Compute the steady state $\bar{\mathbf{F}}^{*(v)}$ on each graph via Eq. (22);
Compute the final learned label matrix by $\bar{\mathbf{F}}^* = \frac{1}{V}\sum_{v=1}^{V} \bar{\mathbf{F}}^{*(v)}$;
Classify every originally unlabeled example to the j-th class via $j = \arg\max_{j' \in \{1, \cdots, c\}} \bar{\mathbf{F}}_{ij'}^*$;
Output: Class labels $y_{l+1}, \cdots, y_{l+u}$;

The above multi-modal teaching and learning process iterates until all the unlabeled examples are propagated, and the resulting label matrix is denoted as $\bar{\mathbf{F}}$. Starting from $\bar{\mathbf{F}}$, the following Eq. (22) is adopted to drive the propagation process of every learner to the steady state, namely

$$\bar{\mathbf{F}}^{*(v)} = (1 - \theta)(\mathbf{I} - \theta \mathbf{P}^{(v)})^{-1} \bar{\mathbf{F}}, \tag{22}$$

where the parameter $\theta = 0.05$. Therefore, the final produced label matrix is $\bar{\mathbf{F}}^* = \frac{1}{V}\sum_{v=1}^{V} \bar{\mathbf{F}}^{*(v)}$, and $\mathbf{x}_i$ is classified into the j-th class that satisfies $j = \arg\max_{j' \in \{1, \cdots, c\}} \bar{\mathbf{F}}_{ij'}^*$. The complete SMMCL algorithm for CL based label propagation is outlined in Algorithm 2.

Experimental Results

In this section, we provide the empirical evaluations of our SMMCL by comparing it with five state-of-the-art methods on four typical image datasets.

Datasets. The four image classification datasets include *Architecture* (Xu et al. 2016) for architecture style recognition, *UIUC* (Li and Li 2007) for sports event classification, *MSRC* (Criminisi 2004) for natural image classification, and *Scene15* (Lazebnik, Schmid, and Ponce 2006) for scene categorization. All the images in the adopted datasets are represented by 72-dimensional Pyramid Histogram Of Gradients (PHOG), 512-dimensional GIST, and 256-dimensional Local Binary Patterns (LBP) features. Therefore, each example is characterized by three different modalities. Note that these feature descriptors are histogram-based and every element in a feature vector falls into $[0, 1]$, so none of them will dominate the learning performance.

Baselines. Five graph-based label propagation algorithms are taken as baselines, which include: 1) Gaussian Field and Harmonic Functions (GFHF) (Zhu and Ghahramani

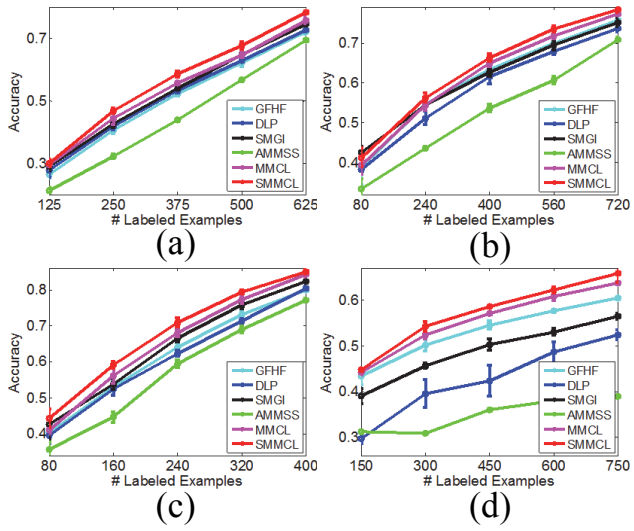

Figure 2: The accuracies of all compared methods on four datasets. (a) is *Architecture*, (b) is *UIUC*, (c) is *MSRC*, and (d) is *Scene15*.

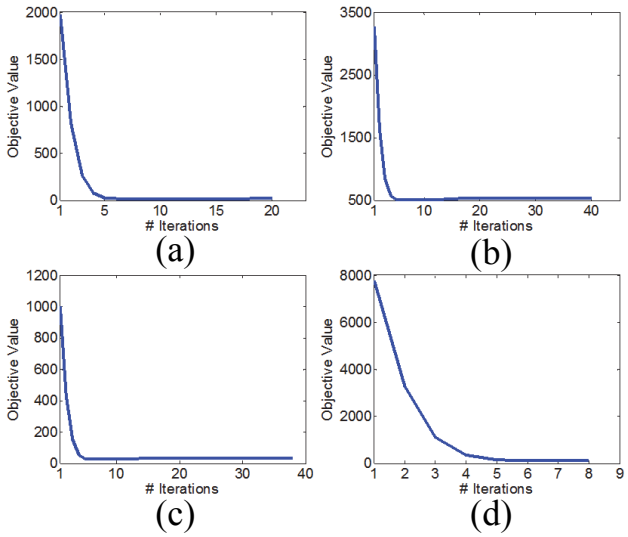

Figure 3: The convergence curves of SMMCL on four datasets. (a) is *Architecture*, (b) is *UIUC*, (c) is *MSRC*, and (d) is *Scene15*.

2002), which is a classical algorithm that serves as the learners in our proposed SMMCL; 2) Dynamic Label Propagation (DLP) (Wang, Tu, and Tsotsos 2013), which is a recently proposed single-modal propagation methodology; 3) Sparse Multiple Graph Integration (SMGI) (Karasuyama and Mamitsuka 2013b) that is a competitive multi-modal graph-based method; 4) Adaptive Multi-Modal Semi-Supervised classifier (AMMSS) (Cai et al. 2013) which is based on multiple graphs and also automatically learns the weight of each modality like our SMMCL; and 5) Multi-Modal Curriculum Learning (MMCL) (Gong et al. 2016b) which is the state-of-the-art CL based algorithm and is very relevant to the proposed method. For the single-modal methods like GFHF and DLP, the GIST, LBP and PHOG feature vectors are directly concatenated into a long feature vector as the inputs.

Experimental settings. For all the datasets, we evaluate the classification accuracies of all compared methods under different sizes of labeled set, and the experiment under each size is implemented five times with different initially labeled examples. The reported accuracies are then obtained by averaging over the outputs of these five independent runs.

For fair comparison, we utilize the graph construction technique in (Karasuyama and Mamitsuka 2013a) and build the identical 10-NN graphs for all comparators in all experiments. The trade-off parameters of our SMMCL are set to $\alpha = 1$ and $\beta = 0.5$. The parameters in SMGI are optimally tuned to $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ via searching the grid $\{0.01, 0.1, 1, 10\}$, and γ and λ in AMMSS are set to 0.5 and 10, respectively. In DLP, we adjust α and λ to 0.05 and 0.1 accordingly as recommended by the authors. Besides, we set $\beta = 10$, $\gamma = 3$ and $\eta = 1.1$ as they lead to the optimal results as revealed by (Gong et al. 2016b).

Results and analyses. The classification accuracies of all compared methods on the four adopted datasets are presented in Fig. 2, which reflects that the performances of all six methods can be improved when the size of labeled examples increases. Besides, we also have several interesting findings regarding the experimental results: firstly, the single-modal methods such as GFHF and DLP generally perform worse than the multi-modal methodologies such as MMCL, SMGI and our SMMCL, which demonstrate that properly combining multi-modal information is better than harnessing only one modality; secondly, the methods based on curriculum learning (*i.e.* MMCL and SMMCL) outperform the other baselines without the curriculum learning scheme, which confirms that learning from simple to difficult can boost the performances of propagation methods; thirdly, among the curriculum learning approaches like MMCL and SMMCL, the state-of-the-art MMCL has already achieved very impressive results, however the proposed SMMCL can still improve the results of MMCL, which validates the superiority of our "soft" multi-modal teaching model to the "hard" one in MMCL; and lastly, our SMMCL consistently leads the incorporated "plain" learner GFHF with a significant margin, which again demonstrates the merits of introducing curriculum learning for simple-to-difficult label propagation.

Illustration of convergence. To show that our SMMCL model can be efficiently solved via ADMM within limited iterations, we plot the objective values of Eq. (5) under different iterations on the above four datasets (see Fig. 3). We see that our algorithm converges quickly and generally terminates within 40 iterations, so the feasibility of employing ADMM for solving Eq. (5) is verified.

Conclusion

This paper proposes a novel multi-modal curriculum learning algorithm for label propagation, which investigates the difficulty of every unlabeled example from multiple modal-

ities and then optimizes the propagation sequence so that the simple examples are classified ahead of the difficult examples. Our method has the following merits: firstly, it comprehensively discovers the commonality among different modalities and meanwhile explicitly exploits their individualities, so the opinions of various teachers are flexibly fused into an unbiased simplest curriculum; secondly, in each learning round the number of selected examples and the weight of each learner for label fusion are adaptively determined based on the level of agreement among all involved teachers; and thirdly, our optimization model can be efficiently solved as every subproblem of Eq. (6) has a closed-form solution. Thorough experimental results demonstrate that our method is superior to several state-of-the-art methodologies on image classification tasks.

Acknowledgments

This research is supported by NSFC, China (No: 61602246). The authors would like to thank the anonymous reviewers for their constructive comments.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proc. International Conference on Machine Learning (ICML)*, 41–48.

Cai, X.; Nie, F.; Cai, W.; and Huang, H. 2013. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *Computer Vision (ICCV), IEEE International Conference on*, 1737–1744.

Chang, X., and Yang, Y. 2016. Semisupervised feature analysis by mining correlations among multiple tasks. *Neural Networks and Learning Systems, IEEE Transactions on*.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 1171–1177.

Chang, X.; Yang, Y.; Long, G.; Zhang, C.; and Hauptmann, A. 2015. Dynamic concept composition for zero-example event detection. In *ACM Multimedia Conference*, 581–590.

Criminisi, A. 2004. Microsoft research cambridge object recognition image dataset. http://research.microsoft.com/en-us/projects/objectclassrecognition/. [Online].

Gong, C.; Tao, D.; Liu, W.; Maybank, S.; Fang, M.; Fu, K.; and Yang, J. 2015. Saliency propagation from simple to difficult. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2531–2539.

Gong, C.; Tao, D.; Liu, W.; Liu, L.; and Yang, J. 2016a. Label propagation via teaching-to-learn and learning-to-teach. *Neural Networks and Learning Systems, IEEE Transactions on*.

Gong, C.; Tao, D.; Maybank, S.; Liu, W.; Kang, G.; and Yang, J. 2016b. Multi-modal curriculum learning for semi-supervised image classification. *Image Processing, IEEE Transactions on* 25(7):3249–3260.

Gong, C.; Tao, D.; Yang, J.; and Liu, W. 2016c. Teaching-to-learn and learning-to-teach for multi-label propagation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Jiang, L.; Meng, D.; Mitamura, T.; and Hauptmann, A. 2014a. Easy samples first: self-paced reranking for zero-example multimedia search. In *ACM Multimedia Conference (ACM MM)*, 547–556.

Jiang, L.; Meng, D.; Yu, S.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014b. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems (NIPS)*, 2078–2086.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. 2015. Self-paced curriculum learning. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Karasuyama, M., and Mamitsuka, H. 2013a. Manifold-based similarity adaptation for label propagation. In *Advances in Neural Information Processing Systems (NIPS)*, 1547–1555.

Karasuyama, M., and Mamitsuka, H. 2013b. Multiple graph label propagation by sparse integration. *Neural Networks and Learning Systems, IEEE Transactions on* 24(12):1999–2012.

Kumar, M.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems (NIPS)*, 1189–1197.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2169–2178.

Lee, Y., and Grauman, K. 2011. Learning the easy things first: Self-paced visual category discovery. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 1721–1728.

Li, L., and Li, F. 2007. What, where and who? classifying events by scene and object recognition. In *Computer Vision (ICCV), IEEE International Conference on*, 1–8.

Lin, Z.; Chen, M.; and Ma, Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv:1009.5055v3* 9.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2012. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(1):171–184.

Pentina, A.; Sharmanska, V.; and Lampert, C. H. 2015. Curriculum learning of multiple tasks. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 5492–5500.

Qiu, H., and Hancock, E. 2007. Clustering and embedding using commute times. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29(11):1873–1890.

Supancic, J., and Ramanan, D. 2013. Self-paced learning for long-term tracking. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2379–2386.

Tang, K.; Ramanathan, V.; Li, F.; and Koller, D. 2012a. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems (NIPS)*, 638–646.

Tang, Y.; Yang, Y.; Yu, B.; and Gao, Y. 2012b. Self-paced

dictionary learning for image classification. In *ACM Multimedia Conference (ACM MM)*, 833–836.

Wang, B.; Tu, Z.; and Tsotsos, J. 2013. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *Computer Vision (ICCV), IEEE International Conference on*, 425–432.

Xu, Z.; Hong, Z.; Zhang, Y.; Wu, J.; Tsoi, A.; and Tao, D. 2016. Multinomial latent logistic regression for image understanding. *Image Processing, IEEE Transactions on* 25(2):973–987.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv:1304.5634*.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view self-paced learning for clustering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 3974–3980.

Zhu, X., and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. International Conference on Machine Learning (ICML)*, 912–919.