Supplementary Material for "Feature Disentangling Machine - A Novel Approach of Feature Selection and Disentangling in Facial Expression Analysis"

Ping Liu[1], Joey Tianyi Zhou[2], Ivor Wai-Hung Tsang[3], Zibo Meng[1], Shizhong Han[1], and Yan Tong[1]

[1]Department of Computer Science, University of South Carolina, USA
[2] Center for Computational Intelligence, Nanyang Technology University, Singapore
[3]Center for Quantum Computation and Intelligent Systems, University of Technology, Australia

1 Introduction

This document serves as a supplementary document for the submission. It illustrates the details of the formulation derivation.

2 Formulation Derivation for Feature Disentanling Machine (FDM)

2.1 Formulation for Feature Disentangling Machine

To simplify the discussion, we only discuss an MTL expression recognition problem considering two expressions, denoted as E_1 and E_2, at the same time. Then, each image sample can be represented by a triplet $\left\{\mathbf{x}_i, y_i^{E_1}, y_i^{E_2}\right\}, i = 1, \cdots, N$ with two expression labels. Specifically, if one of the target expressions, e.g., E_1, is activated in the sample, then $y_i^{E_1} = 1$ and $y_i^{E_2} = -1$, and vice versa; while if neither E_1 nor E_2 is activated, both the expression labels are set to -1. $y_i^{E_1}$ and $y_i^{E_2}$ cannot be 1 both since for each image there is one expression label.

In order to select *expression-specific features*, two *expression-specific feature selection* vectors denoted as $\mathbf{d}^{E_1}$ and $\mathbf{d}^{E_2}$ are introduced for tasks E_1 and E_2, respectively. In addition, a *common feature selection* vector denoted as $\mathbf{d}^{E_c}$ is used to select common features that are effective and shared in recognizing all expressions.

Therefore, the objective function in proposed Feature Disentangling Machine can be extended to recognizing both expressions simultaneously in an MTL framework as follows:

$$\min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\in\mathcal{D}\}}\ \min_{\{\mathbf{w}^{E_1},\mathbf{w}^{E_2},\boldsymbol{\epsilon}^{E_1},\boldsymbol{\epsilon}^{E_2},\rho_1,\rho_2\}} \frac{1}{2}\left(\|\mathbf{w}^{E_1}\|_2^2+\|\mathbf{w}^{E_2}\|_2^2\right)+\frac{\gamma}{2}\sum_{i=1}^{N}\left[(\epsilon_i^{E_1})^2+(\epsilon_i^{E_2})^2\right]-(\rho_1+\rho_2)$$

$$\begin{aligned} s.t. \quad & y_i^{E_1}(\mathbf{w}^{E_1})^T\left[\mathbf{x}_i\circ\left(\mathbf{d}^{E_1}+\mathbf{d}^{E_c}\right)\right]\geq\rho_1-\epsilon_i^{E_1}, i=1,\cdots,N,\\ & y_i^{E_2}(\mathbf{w}^{E_2})^T\left[\mathbf{x}_i\circ\left(\mathbf{d}^{E_2}+\mathbf{d}^{E_c}\right)\right]\geq\rho_2-\epsilon_i^{E_2}, i=1,\cdots,N.\\ & \sum_{j=1}^{m} d_j^{E_1}\leq\tau_1 \qquad \sum_{j=1}^{m} d_j^{E_2}\leq\tau_2 \qquad \sum_{j=1}^{m} d_j^{E_c}\leq\tau_c \qquad d_j^{E_1},d_j^{E_2},d_j^{E_c}\in\{0,1\}\\ & d_j^{E_1}+d_j^{E_2}+d_j^{E_c}\leq 1 \qquad j=1,\cdots,m. \end{aligned} \tag{1}$$

where the sparsity of features is controlled in the three feature selection vectors (i.e., $\mathbf{d}^{E_1}$, $\mathbf{d}^{E_2}$, and $\mathbf{d}^{E_c}$) by three parameters τ_1, τ_2, and τ_c.

In the subsequent discussion, we will present the details of the algorithm to solve the FDM.

2.2 Algorithm for Solving Feature Disentangling Machine

First we transform the inner minimization problem into the dual of SVM by following procedures:

$$\begin{aligned} & L_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}}(\mathbf{w}^{E_1},\mathbf{w}^{E_2},\boldsymbol{\epsilon}^{E_1},\boldsymbol{\epsilon}^{E_2},\boldsymbol{\alpha},\boldsymbol{\beta})\\ &=\frac{1}{2}\left(\|\mathbf{w}^{E_1}\|_2^2+\|\mathbf{w}^{E_2}\|_2^2\right)+\frac{\gamma}{2}\left[\sum_{i=1}^{N}(\epsilon_i^{E_1})^2+\sum_{i=1}^{N}(\epsilon_i^{E_2})^2\right]-(\rho_1+\rho_2)\\ &-\sum_i\alpha_i\left\{y_i^{E_1}(\mathbf{w}^{E_1})^T\left[(\mathbf{d}^{E_1}+\mathbf{d}^{E_c})\circ\mathbf{x}_i\right]+\epsilon_i^{E_1}-\rho_1\right\}\\ &-\sum_i\beta_i\left\{y_i^{E_2}(\mathbf{w}^{E_2})^T\left[(\mathbf{d}^{E_2}+\mathbf{d}^{E_c})\circ\mathbf{x}_i\right]+\epsilon_i^{E_2}-\rho_2\right\} \end{aligned} \tag{2}$$

Take the partial derivative of $L_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}}(\mathbf{w}^{E_1},\mathbf{w}^{E_2},\boldsymbol{\epsilon}^{E_1},\boldsymbol{\epsilon}^{E_2},\boldsymbol{\alpha},\boldsymbol{\beta})$ w.r.t $\mathbf{w}^{E_1}$, $\mathbf{w}^{E_2}$, $\boldsymbol{\epsilon}^{E_1}$, $\boldsymbol{\epsilon}^{E_2}$, ρ_1, ρ_2 and set the values to 0. We can have

$$
\begin{aligned}
&\mathbf{w}^{E_1} = \sum_{i=1}^{N} \alpha_i y_i^{E_1} \left[(\mathbf{d}^{E_1} + \mathbf{d}^{E_c}) \circ \mathbf{x}_i \right]; \\
&\mathbf{w}^{E_2} = \sum_{i=1}^{N} \beta_i y_i^{E_2} \left[(\mathbf{d}^{E_2} + \mathbf{d}^{E_c}) \circ \mathbf{x}_i \right]; \\
&\gamma \epsilon_i^{E_1} = \alpha_i \\
&\gamma \epsilon_i^{E_2} = \beta_i \\
&\sum_{i=1}^{N} \alpha_i = 1 \\
&\sum_{i=1}^{N} \beta_i = 1 \\
&i = 1, \cdots, N
\end{aligned}
\tag{3}
$$

Substituting Eq. 3 to Eq. 2 (the original Lagrangian function), we have the following formulation:

$$
\begin{aligned}
&L_{\{\mathbf{d}^{E_1}, \mathbf{d}^{E_2}, \mathbf{d}^{E_c}\}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \\
&-\frac{1}{2} \left\| \sum_i \alpha_i y_i^{E_1} \left[(\mathbf{d}^{E_1} + \mathbf{d}^{E_c}) \circ \mathbf{x}_i \right] \right\|^2 \\
&-\frac{1}{2} \left\| \sum_i \beta_i y_i^{E_2} \left[(\mathbf{d}^{E_2} + \mathbf{d}^{E_c}) \circ \mathbf{x}_i \right] \right\|^2 \\
&-\frac{1}{2\gamma} \boldsymbol{\alpha}^T \boldsymbol{\alpha} - \frac{1}{2\gamma} \boldsymbol{\beta}^T \boldsymbol{\beta}
\end{aligned}
\tag{4}
$$

Consequently the original inner problem (Eq. 1) is transformed into its dual formulation, in which the solution can be found by solving the corresponding dual problem as:

$$\min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} L_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta}) =$$

$$\min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} -\frac{1}{2}\left\|\sum_i \alpha_i y_i^{E_1}\left[\left(\mathbf{d}^{E_1}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\alpha}^T\boldsymbol{\alpha}$$

$$-\frac{1}{2}\left\|\sum_i \beta_i y_i^{E_2}\left[\left(\mathbf{d}^{E_2}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{5}$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i = 1,\quad \sum_{i=1}^{N}\beta_i = 1,\quad \alpha_i > 0,\quad \beta_i > 0,\qquad for\ i = 1,\cdots,N,$$

$$\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}\in\mathcal{D},$$

$$where\qquad \mathcal{D} = \{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}|\sum_{j=1}^{m} d_j^{E_1}\le\tau_1,\sum_{j=1}^{m} d_j^{E_2}\le\tau_2,\sum_{j=1}^{m} d_j^{E_c}\le\tau_c,$$

$$d_j^{E_1}+d_j^{E_2}+d_j^{E_c}\le 1,\qquad d_j^{E_1},d_j^{E_2},d_j^{E_c}\in\{0,1\},\qquad for\ j=1,\cdots,m\}$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are dual variable vectors for the inequality constraints in the inner minimization problem (Eq. 1).

The saddle point problem (5) can be lower bounded by:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} L_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta}) =$$

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} -\frac{1}{2}\left\|\sum_i \alpha_i y_i^{E_1}\left[\left(\mathbf{d}^{E_1}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\alpha}^T\boldsymbol{\alpha}$$

$$-\frac{1}{2}\left\|\sum_i \beta_i y_i^{E_2}\left[\left(\mathbf{d}^{E_2}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{6}$$

$$s.t.\ \sum_{i=1}^{n}\alpha_i = 1,\ \sum_{i=1}^{n}\beta_i = 1,\ \alpha_i > 0,\ \beta_i > 0,\ for\ i = 1,...,N,\ \{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}\in\mathcal{D}$$

By bringing an additional variable θ, the above optimization problem (Eq. 6) becomes:

$$\max_{\theta,\boldsymbol{\alpha},\boldsymbol{\beta}} -\theta : \theta \ge -L_{\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta}),\qquad \forall\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}\in\mathcal{D} \tag{7}$$

which is a convex Quadratically Constrained Quadratic Programming (QCQP) problem.

Define $\mu_t \ge 0$ as the dual variable for each constraint in Eq. 7, the Lagrangian of Eq. 7 can be written as:

$$S(\theta,\boldsymbol{\mu}) = -\theta + \sum_t \mu_t\left[\theta + L_{\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta})\right] \tag{8}$$

Setting derivative of Eq. 8 w.r.t θ to zero, we have:

$$\sum_t \mu_t = 1 \tag{9}$$

Define $\mathcal{M} = \{\boldsymbol{\mu} | \sum_t \mu_t = 1, \mu_t \geq 0\}$ and let $\mathcal{M}$ be the domain of $\boldsymbol{\mu}$, the Lagrangian $S(\theta, \boldsymbol{\mu})$ (Eq. 8) can be rewritten as:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{\mu}\in\mathcal{M}} S(\theta, \boldsymbol{\mu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_t \mu_t L_{\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{10}$$

Since the objective function $\sum_t \mu_t L_{\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is concave in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and is convex in $\boldsymbol{\mu}$, then

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{\mu}\in\mathcal{M}} \sum_t \mu_t L_{\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \min_{\boldsymbol{\mu}\in\mathcal{M}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} -\frac{1}{2}\left(\boldsymbol{\alpha} \circ \mathbf{y}^{E_1}\right)^T \left(\sum_t \mu_t X_t^{E_1} {X_t^{E_1}}^T + \frac{1}{\gamma}\mathcal{I}\right)\left(\boldsymbol{\alpha} \circ \mathbf{y}^{E_1}\right) \\
&\qquad - \frac{1}{2}\left(\boldsymbol{\beta} \circ \mathbf{y}^{E_2}\right)^T \left(\sum_t \mu_t X_t^{E_2} {X_t^{E_2}}^T + \frac{1}{\gamma}\mathcal{I}\right)\left(\boldsymbol{\beta} \circ \mathbf{y}^{E_2}\right)
\end{aligned} \tag{11}$$

$$\begin{aligned}
\textit{where} \quad & X_t^{E_1} = \left[\mathbf{x}_1 \circ \left(\mathbf{d}_t^{E_1} + \mathbf{d}_t^{E_c}\right), \cdots, \mathbf{x}_N \circ \left(\mathbf{d}_t^{E_1} + \mathbf{d}_t^{E_c}\right)\right]^T \\
& X_t^{E_2} = \left[\mathbf{x}_1 \circ \left(\mathbf{d}_t^{E_2} + \mathbf{d}_t^{E_c}\right), ..., \mathbf{x}_N \circ \left(\mathbf{d}_t^{E_2} + \mathbf{d}_t^{E_c}\right)\right]^T \\
& \mathcal{M} = \{\boldsymbol{\mu} | \sum \mu_t = 1, \mu_t \geq 0\}; \{\mathbf{d}_t^{E_1}, \mathbf{d}_t^{E_2}, \mathbf{d}_t^{E_c}\} \in \mathcal{D}
\end{aligned}$$

where $\mathcal{I}$ represents an identity matrix.

The details to solve the Eq. 11 can be found in the paper submitted.