**Multi-label Classification Setup**

Which tags are associated with this picture?

\[ Y = \{ \text{building, taipei 101, day, night view, skyscraper, fireworks, sun, face, firework, sun face, taipei world financial center, unsatellite, etc.} \} \]

- Given: N examples \( \{ \mathbf{x}_i \in \mathbb{R}^D, y_i \subseteq \{1, 2, \ldots, K \}\}_{i=1}^N \)
- Goal: classifier \( g(y) \) that closely predicts the label-set \( Y \) associated with some unseen inputs \( x \), presumably by exploiting hidden relations between labels, e.g.
  - taipei 101 & taipei world financial center highly correlated
  - skyscraper subset of building
  - day view & night view disjoint

**Label Space Dimension Reduction**

\[ Y \subseteq \{1, 2, \ldots, K\} \text{ equivalent to } y \in \{0, 1\}^K \]

- feature space dimension reduction: compress \( x \) to remove irrelevant, redundant (possibly related), or noisy information, and achieve better efficiency & performance
  - principal component analysis (PCA): linearly project \( x \) to \( W \mathbf{x} \) with minimum projection error
  - canonical correlation analysis (CCA): linearly project \( x \) to \( W \mathbf{x} \) in order to maximize correlation with \( Y \)
- label space dimension reduction: analogously, but compress \( y \) instead
  1. compress: transform \( \{(x_i, y_i)\} \rightarrow \{(x_i, t_i)\} \) with \( t_i = \text{compress}(y_i) \in \{0, 1\} \) and \( M < K \)
  2. learn: train some \( f(x) \) from \( \{(x_i, t_i)\} \)
  3. decompress: \( g(x) = \text{decompress}(f(x)) \)

**Conditional Principal Label Space Transformation**

- idea 1: exploit dual role of CCA to be feature-aware
  - project \( x \) to \( W \mathbf{x} \) in order to maximize correlation with some \( V \mathbf{y} \)
  - project \( y \) to \( V \mathbf{y} \) in order to minimize difference to some \( W \mathbf{x} \)
  - proposed OCCA: \( \min W, V \{ |XW^T − ZV^T|_2^2 \} \), s.t. \( V \mathbf{y} \) = I
    - project to easiest-by-linear-regression directions
  - idea 2: keep benefits of PLST for compression
    - existing PLST: \( \min |Y − VV^T\mathbf{y}|_2 \), s.t. \( V \mathbf{y} \) = I
    - project to most representative directions
  - proposed algorithm: conditional principal label space transformation (CPLST)
    \[ \min W \{ |XW^T − YV^T|_2^2 + |Y − VV^T\mathbf{y}|_2^2 \} \text{ s.t. } V \mathbf{y} \equiv I \]
    - theoretical guarantee (Tai and Lin, NC 2012): when using linear regression as \( r \),
      - learning error \( \geq \) compression error
  - algorithmic simplicity: closed-form optimal \( V \) contains top eigenvectors of \( XX^T \)

- compressive sensing (Bin et al., NIPS 2009): linearly project \( y \) to \( t = \text{compress}(y) \) with random \( v \) (for coherence
- principal label space transformation (PLST): Tai and Lin, NC 2012): linearly project \( y \) to \( t = \text{compress}(y) \) with minimum projection error (sibling of PCA)

**Feature-Aware Label Space Dimension Reduction**

- feature space dimension reduction: \( \min |XW^T − ZV^T|_2 \) \text{ s.t. } \mathbf{y} = I
  - supervised (using \( y \))
    - PCA, locally linear embedding, etc.
    - CCA, sliced inverse regression, etc.
  - unsupervised (not using \( y \))
    - supervised generally better for learning from \( \text{compr}(x) \)

- label space dimension reduction
  - feature-aware (using \( x \))
    - CPLST, compressive sensing, etc.
  - feature-unaware (not using \( x \))
    - PLST, conditional principal label space transformation

- can we improve PLST by feature-aware label space dimension reduction?

**Experimental Results**

- OCCA: optimize learning error, but worst in compression error
- PLST: optimize compression error, but worst in learning error
- CPLST: optimize learning+compression error, and hence best hamming loss on benchmark data sets

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CPLST vs. PLST</th>
<th>PLST vs. PLST</th>
<th>OCCA vs. PLST</th>
<th>CPLST vs. OCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>M = 20/K</td>
<td>2 win, 5 similar</td>
<td>2 win, 6 similar</td>
<td>5 win, 1 lose, 2 similar</td>
<td>CPLST consistently better than or similar to PLST across data &amp; algorithms</td>
</tr>
</tbody>
</table>

**Summary**

Conditional Principal Label Space Transformation, which
- projects to conditional principal directions by combining ideas behind CCA (feature-aware) and PLST (optimal compression).
- can be kernelized for exploiting feature power.
- achieves better/similar practical performance consistently when compared with the readily-strong PLST.

---

**Feature-aware Label Space Dimension Reduction for Multi-label Classification**

Yao-Nan Chen (r99922008@cseie.ntu.edu.tw) and Hsuan-Tien Lin (htlin@cseie.ntu.edu.tw)
Department of Computer Science and Information Engineering, National Taiwan University