1. Problem

- Introduction:
  - Data is often a collection of modalities
  - Images and captions, audio and video, etc.
  - Need to fuse the data into a shared representation that captures the “concept”

- Applications:
  - Image Retrieval
  - Image Classification

2. MIR Flickr Dataset

- Images and tags collected from Flickr
- 1 Million Images and their Tags
- 25K have been annotated for 38 classes
- Each image can belong to several classes
- Example classes people, female, tree, and water

<table>
<thead>
<tr>
<th>Tag</th>
<th>#Images</th>
<th>Tag</th>
<th>#Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td>645</td>
<td>cloudy</td>
<td>376</td>
</tr>
<tr>
<td>water</td>
<td>141</td>
<td>ocean</td>
<td>1362</td>
</tr>
<tr>
<td>portrait</td>
<td>423</td>
<td>sea</td>
<td>301</td>
</tr>
<tr>
<td>sport</td>
<td>421</td>
<td>river</td>
<td>300</td>
</tr>
<tr>
<td>nature</td>
<td>556</td>
<td>palm</td>
<td>262</td>
</tr>
<tr>
<td>animal</td>
<td>385</td>
<td>crowd</td>
<td>256</td>
</tr>
<tr>
<td>cloud</td>
<td>209</td>
<td>fog</td>
<td>220</td>
</tr>
<tr>
<td>flower</td>
<td>351</td>
<td>bird</td>
<td>215</td>
</tr>
<tr>
<td>beach</td>
<td>407</td>
<td>dog</td>
<td>214</td>
</tr>
<tr>
<td>landscape</td>
<td>395</td>
<td>cat</td>
<td>212</td>
</tr>
<tr>
<td>street</td>
<td>303</td>
<td>lake</td>
<td>109</td>
</tr>
<tr>
<td>art</td>
<td>373</td>
<td>building</td>
<td>108</td>
</tr>
<tr>
<td>architecture</td>
<td>354</td>
<td>river</td>
<td>175</td>
</tr>
<tr>
<td>graffiti</td>
<td>335</td>
<td>baby</td>
<td>167</td>
</tr>
<tr>
<td>house</td>
<td>133</td>
<td>animal</td>
<td>164</td>
</tr>
</tbody>
</table>

Average number of tags is 5.15

3. Theory

- **Autoencoders**: The aim of an autoencoder is to learn a (encoding) for a set of data

  **Input**: $X \in [0,1]^d$  
  **Output**: $Z \in [0,1]^d$

  Maps $X$ to $Y$ through a mapping $y = s(Wx + b)$

  Decoder maps $Y$ into a reconstruction $Z$ through $z = s(W'y + b')$

  Parameters $W$ and $W'$ are optimized such that the average reconstruction error $L(y,z) = \frac{1}{d} \sum_{i=1}^{d} (y_i \log z_i + (1 - y_i) \log (1 - z_i))$ is minimized

- **Denoising Autoencoders**: Trains the autoencoder to reconstruct the input from a corrupted version of the input

  **Input**: $X \in [0,1]^d$, Corrupted input: $X'$

  $X'$ is mapped to $Y$ through $y = f_\theta(x') = s(Wx' + b)$

  We reconstruct $z = g_\theta(y)$

  Parameters $\theta$ and $\theta'$ are trained to minimize the reconstruction error, $L(x,z) = \|x - z\|^2$

4. Results

**Image Retrieval**

- Given: Tags  
- Output: Similar images

**Multimodal Stacked Denoising Autoencoders**

- Pretrained using 975K unlabeled Flickr images and tags
- Shared representation used for classification and retrieval

**Retrieved Images**

- NYC, New York, Taxi, Times Square, Night

**Method**  
- Random  
- LDA  
- SVM  
- DBM  
- SDA (Proposed)

**Average reconstruction error**

- Random: 0.124  
- LDA: 0.492  
- SVM: 0.475  
- DBM: 0.609  
- SDA (Proposed): 0.533

**Sky**

- Airplane, cloud