Ancient Document Layout Analysis: Autoencoders meet Sparse Coding

Homa Davoudi, Marco Fiorucci, Arianna Traviglia

Center for Cultural Heritage Technology (CCHT), Istituto Italiano di Tecnologia (IIT)
Venice, Italy
Historical Document Layout Analysis

- Segmenting into homogeneous regions:
  - Blocks of text, side notes, drawings, tables, etc.
- Key preprocessing step in various applications
- An open problem for historical documents
  - No structured text arrangement
  - High degradation

Deep Neural Networks for Doc layout Analysis

- Two main Categories:
  - Image to image mapping
  - Representation learning based methods

A novel **unsupervised sparse representation learning** method
Propose DLA Pipeline

- Extract fixed size patches
- Compute Sparse representation vectors
- Classify each pixel by a feed-forward network
Neural Sparse Coding

- Classical sparse coding:
  - Restricted to linear combination of sparse features and dictionary atoms

- DNN based sparse coding:
  - Train sparse encoder in a supervised way
  - Model iterative optimization steps

**Proposed method:** Encoder and Sparse representation trained simultaneously without unfolding
Proposed Sparse Representation Learning

- Model Architecture
  - Encoder-Decoder with sparse latent Variables

![Diagram of Encoder-Decoder Model](image)

- CNN Encoder
- S: Soft shrinkage function
- Sparse code
- CNN Decoder
- \( x \) -> \( h_e \) -> \( h_d \) -> \( \tilde{x} \)

Dictionary

Forward pass

Backward pass
Proposed Sparse Representation Learning

- Model Architecture

Classical Sparse Coding:

\[
\min_D \frac{1}{T} \sum_{i=1}^{T} \min_{h^{(i)}} \left( \frac{1}{2} \|x^{(i)} - Dh^{(i)}\|_2^2 + \lambda \|h^{(i)}\|_1 \right)
\]

Proposed architecture:

\[
\min_D \frac{1}{T} \sum_{i=1}^{T} \min_{w} \left( \frac{1}{2} \|x^{(i)} - G(D, S(F(x^{(i)})))\|_2^2 + \lambda \|h^{(i)}\|_1 \right)
\]
Proposed Sparse Representation Learning

- Training
  - Dictionary learning:
    - Main strength of sparse coding is in encoding algorithm (not the learned dictionary)[1]
    - We adapt the dictionary learned by VQ-VAE [2]

\[ x \xrightarrow{\text{CNN}} h_e \xrightarrow{\text{NN dictionary look-up}} h_d \xrightarrow{\text{CNN}} \tilde{x} \]

Proposed Sparse Representation Learning

- **Training**
  - **Encoder Training:** Inspired by the ISTA algorithm

\[
\min_D \frac{1}{T} \sum_{i=1}^{T} \min_{h^{(i)}} \frac{1}{2} \|x^{(i)} - Dh^{(i)}\|_2^2 + \lambda \|h^{(i)}\|_1
\]

\[
\begin{align*}
(1) & \quad h^{temp}_{t+1} = h_t - \alpha \nabla \|x - Dh_t\|_2^2 \\
(2) & \quad h^{t+1} = \text{shrink}(h^{temp}_{t+1}, \lambda \alpha)
\end{align*}
\]

minimize $\ell_2$ distance

Apply $\ell_1$ norm

\[
\begin{align*}
(1) & \quad w^{temp}_{t+1} = w_t - \alpha \nabla \|x - G(D, h_{w_t}(x))\|_2^2 \\
(2) & \quad h_{w_{t+1}}(x) = \text{shrink}(F_w^{temp}(x), \lambda \alpha)
\end{align*}
\]

Backward pass

Forward pass
Experiments and Results

- **DIVA-HisDB dataset**
  - A collection of three medieval manuscripts
    - 40 pages for each: 20 Train, 10 Validation, 10 Test
    - background, main text, comments and decoration
  - Pixel-level annotation
  - image patches of 64 x 64
## Experiments and Results

### Comparison Results

<table>
<thead>
<tr>
<th></th>
<th>CB55</th>
<th></th>
<th></th>
<th>CSG18</th>
<th></th>
<th></th>
<th>CSG863</th>
<th></th>
<th>Overall</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc(%)</td>
<td>IU(%)</td>
<td>F1(%)</td>
<td>Acc(%)</td>
<td>IU(%)</td>
<td>F1(%)</td>
<td>Acc(%)</td>
<td>IU(%)</td>
<td>F1(%)</td>
<td>Acc(%)</td>
<td>IU(%)</td>
<td>F1(%)</td>
</tr>
<tr>
<td>Sparse encoding</td>
<td>98.35</td>
<td>79.81</td>
<td>72.14</td>
<td>92.96</td>
<td>77.82</td>
<td>59.30</td>
<td>97.74</td>
<td>73.21</td>
<td>58.72</td>
<td>96.35</td>
<td>76.94</td>
<td>63.38</td>
</tr>
<tr>
<td>VQ-VAE</td>
<td>96.11</td>
<td>66.38</td>
<td>58.35</td>
<td>96.38</td>
<td>69.70</td>
<td>57.23</td>
<td>97.10</td>
<td>68.73</td>
<td>53.69</td>
<td>96.52</td>
<td>68.27</td>
<td>56.42</td>
</tr>
<tr>
<td>CAE [3]</td>
<td>94.31</td>
<td>–</td>
<td>–</td>
<td>95.36</td>
<td>–</td>
<td>–</td>
<td>96.98</td>
<td>–</td>
<td>–</td>
<td>95.55</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Thanks for Your Attention

Waiting to meet you virtually and discuss more!