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)
Email: {homa.davoudi, marco.fiorucci, arianna.traviglia} @iit.it

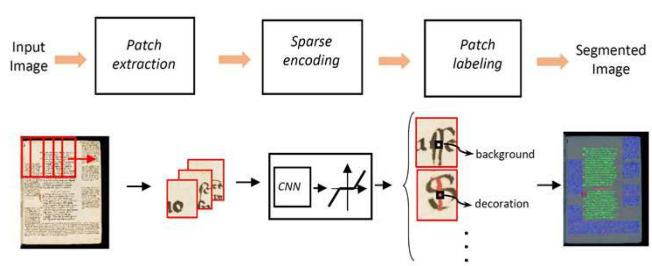
ISTITUTO ITALIANO DI TECNOLOGIA
CENTER FOR CULTURAL HERITAGE TECHNOLOGY

Historical Document Layout Analysis

- Segmenting image into homogeneous regions:
  - Blocks of text, side notes, drawings, tables, etc.
- Key preprocessing step in various applications
- Still an open problem for **historical documents**:
  - Usually lacking a structured text arrangement
  - High degradation
- Deep Neural Networks for Doc layout Analysis:
  - pixel classification methods
  - feature learning based methods

➡ A novel **unsupervised** representation learning method for DLA:
  - Based on the **sparse representation** of image patches

DLA pipeline



- Fixed size patches extracted.
- Sparse representation vector computed for each extracted patch
- A feed-forward network is trained to classify each pixel

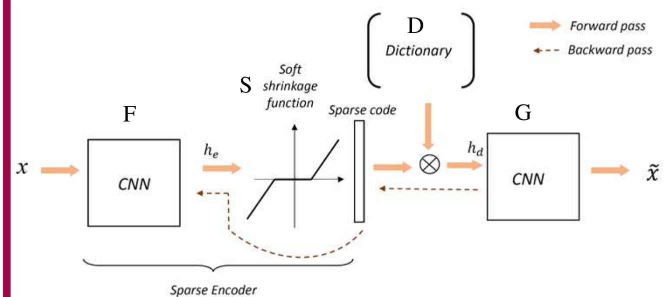
Neural Sparse Coding

- **Classical sparse coding** :
  - restricted to the linear combination of sparse feature vector and dictionary atoms
- Recent **DNN based sparse coding**:
  - Train the encoder in a supervised way.
  - modelling the iterative optimization steps via unfolding a neural network.

➡ **Our method:**
  - Encoder and sparse representation are trained simultaneously in an end-to-end fashion.

Sparse Representation Learning

- Architecture
  - Encoder-Decoder with sparse latent Variables



Classical Sparse Coding: $$\min_D \frac{1}{T}\sum_{i=1}^{T} \min_{h^{(i)}} \frac{1}{2}\left\|x^{(i)} - Dh^{(i)}\right\|_2^2 + \lambda\left\|h^{(i)}\right\|_1$$

Proposed architecture:

$$\min_D \frac{1}{T}\sum_{i=1}^{T} \min_w \frac{1}{2}\left\|x^{(i)} - G\left(D, S(F(x^{(i)}))\right)\right\|_2^2 + \lambda\left\|h^{(i)}\right\|_1$$

- 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]
  - Encoder Training:

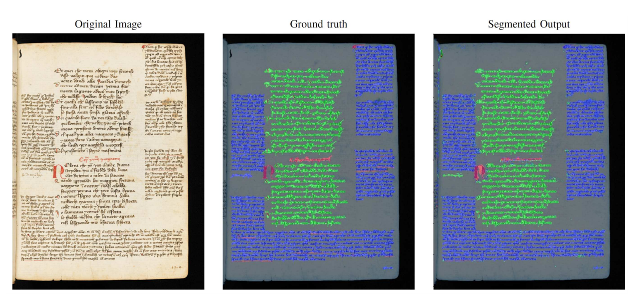
    Inspired by the ISTA algorithm, our encoder network is trained in an iterative way

$$\begin{cases} (1) \quad w_{t+1}^{temp} = w_t - \alpha\nabla\left\|x - G\left(D, h_{w_t}(x)\right)\right\|_2^2 & \text{Backward pass} \\ (2) \quad h_{w_{t+1}}(x) = shrink(F_{w_{t+1}^{temp}}(x), \lambda\alpha) & \text{Forward pass} \end{cases}$$

Experiments and Results

- Experiments on DIVA-HisDB dataset [3]

	CB55			CSG18			CSG863			Overall		
	Acc(%)	IU(%)	F1(%)	Acc(%)	IU(%)	F1(%)	Acc(%)	IU(%)	F1(%)	Acc(%)	IU(%)	F1(%)
Sparse encoding	**98.35**	**79.81**	**72.14**	92.96	**77.82**	**59.30**	**97.74**	**73.21**	**58.72**	96.35	**76.94**	**63.38**
VQ-VAE	96.11	66.38	58.35	**96.38**	69.70	57.23	97.10	68.73	53.69	**96.52**	68.27	56.42
1-layer CNN [4]	60	48	_	53	42	_	57	45	_	56/7	45	_
CAE [3]	94.31	_	_	95.36	_	_	96.98	_	_	95.55	_	_



References

[1] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in ICML, 2011.

[2] A. van den Oord, O. Vinyals et al., "Neural discrete representation learning," in Neurips 2017.

[3] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, "Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts," in ICFHR 2016.

[4] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in ICDAR 2017.