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

ISTITUTO ITALIANO
DI TECNOLOGIA
CENTER FOR CULTURAL
HERITAGE TECHNOLOGY

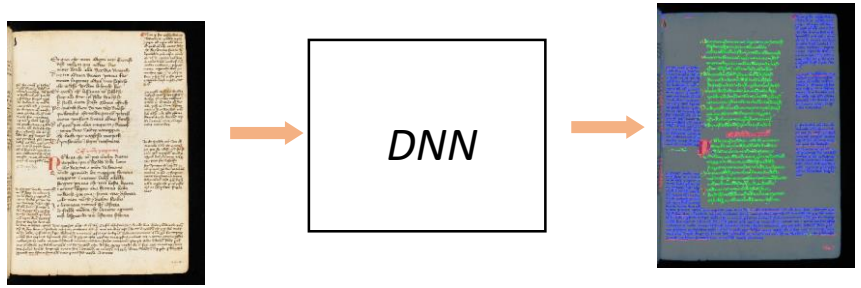
Historical Document Layout Analysis

- 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- An open problem for historical documents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

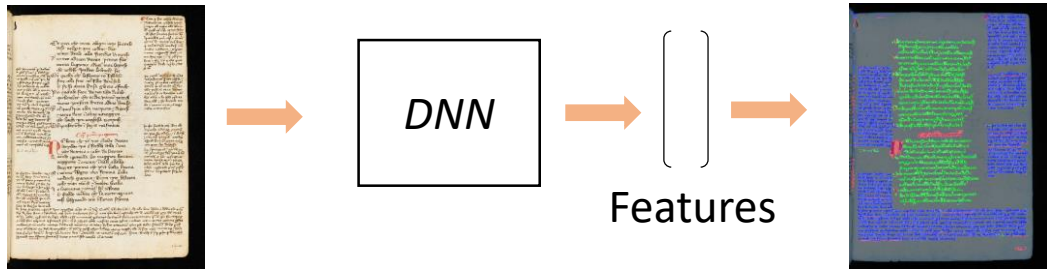
[G. M. Binmakhashen and S. A. Mahmoud, “Document layout analysis: A comprehensive survey,” ACM Comput. Surv., vol. 52, no. 6, 2019.]

Deep Neural Networks for Doc layout Analysis

- Two main Categories:
  - Image to image mapping



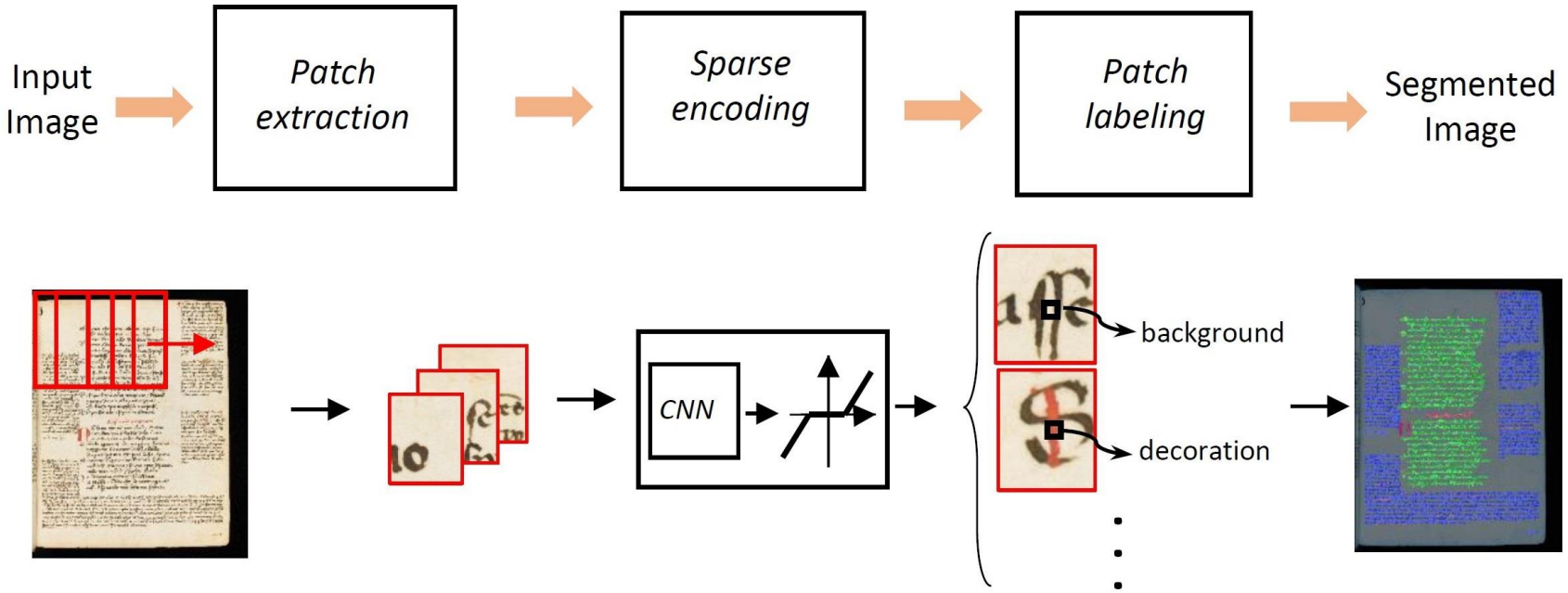
  - Representation learning based methods



- A novel **unsupervised sparse representation learning** method

Propose DLA Pipeline

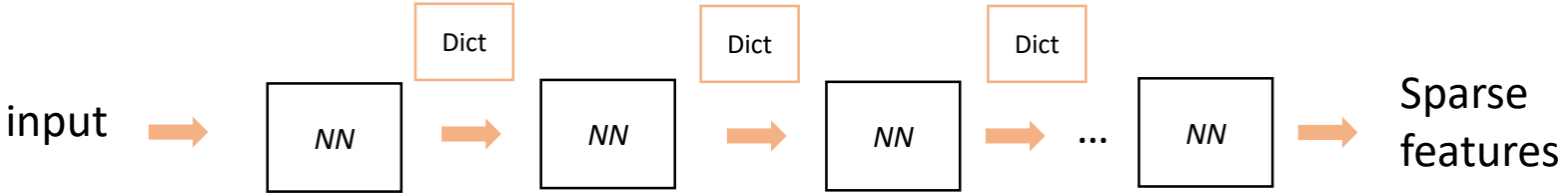
- Extract fixed size patches
- Compute Sparse representation vectors
- Classify each pixel by a feed-forward network



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

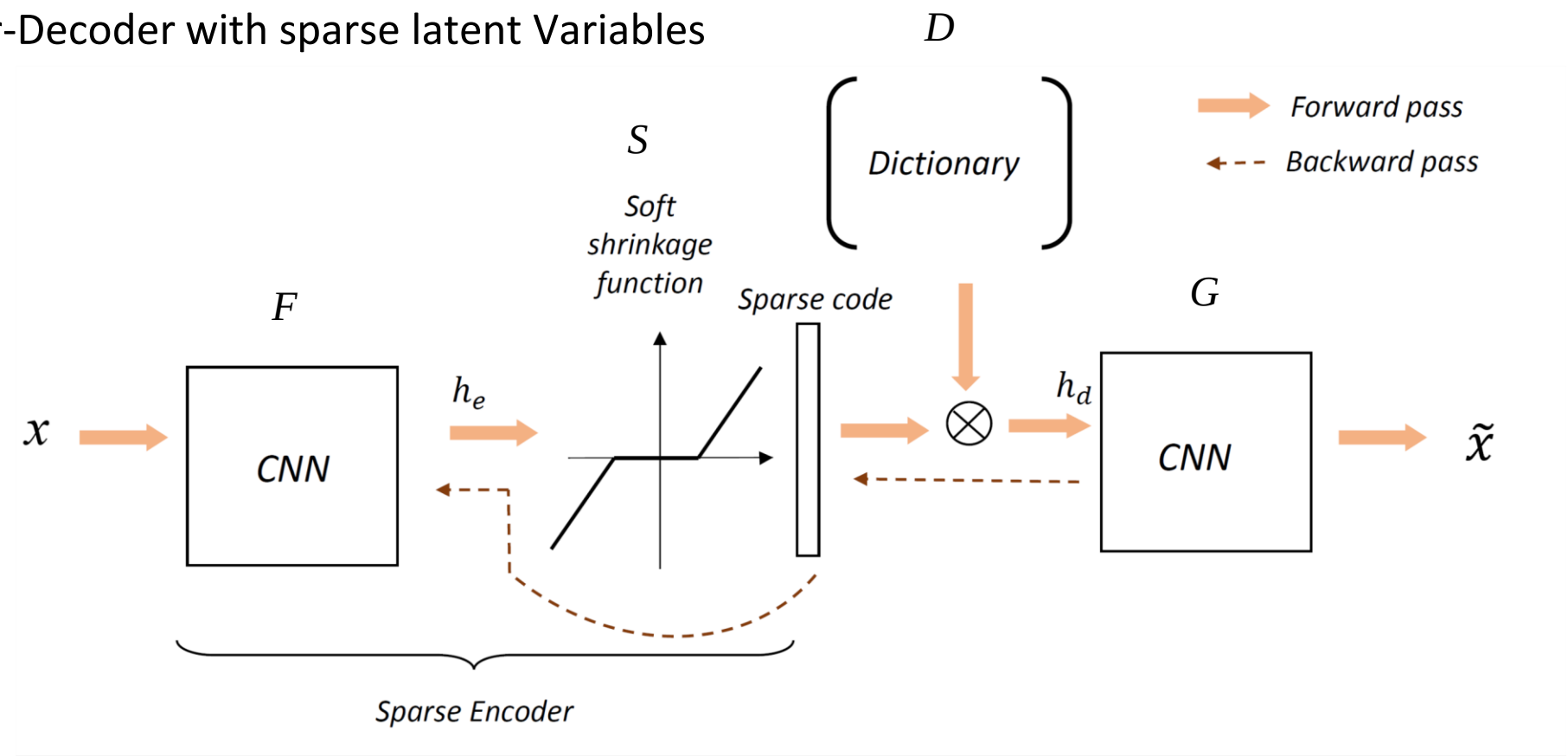
  - model iterative optimization steps



Proposed method: Encoder and Sparse representation trained simultaneously without unfolding

Proposed Sparse Representation Learning

- Model Architecture
  - Encoder-Decoder with sparse latent Variables



Proposed Sparse Representation Learning

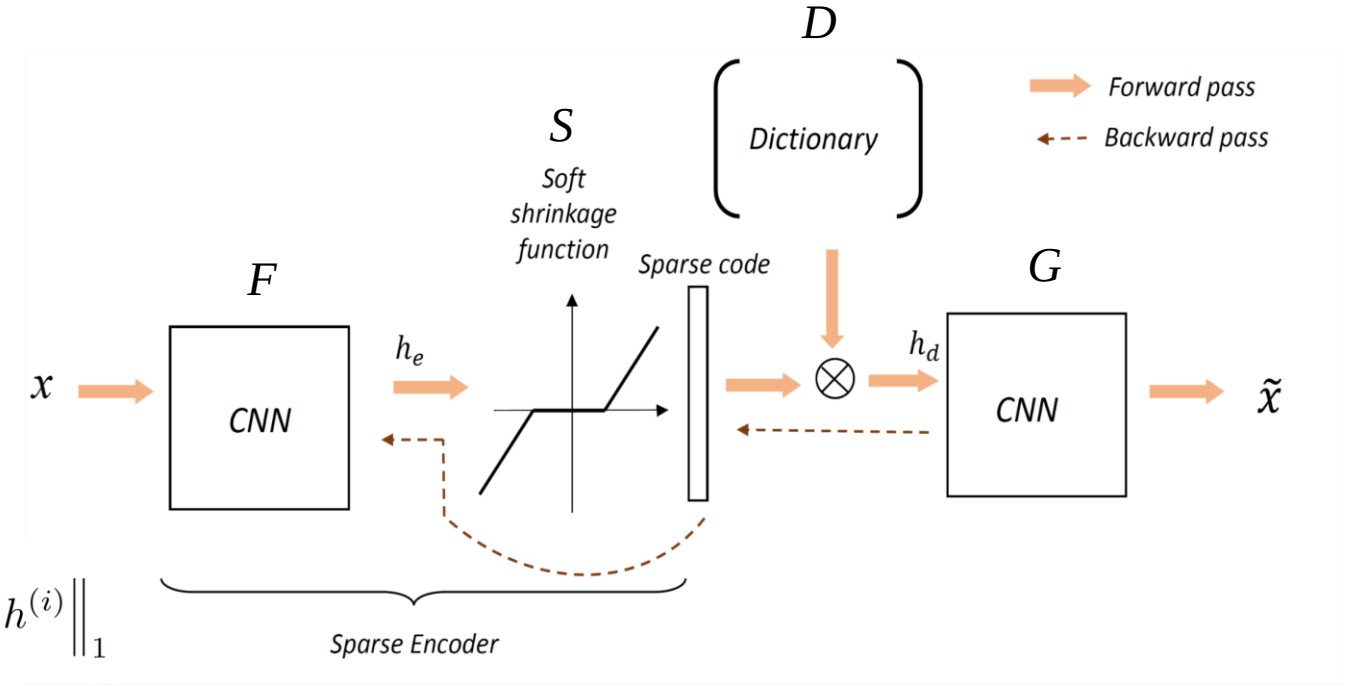
- Model Architecture

Classical Sparse Coding:

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

Proposed architecture:

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



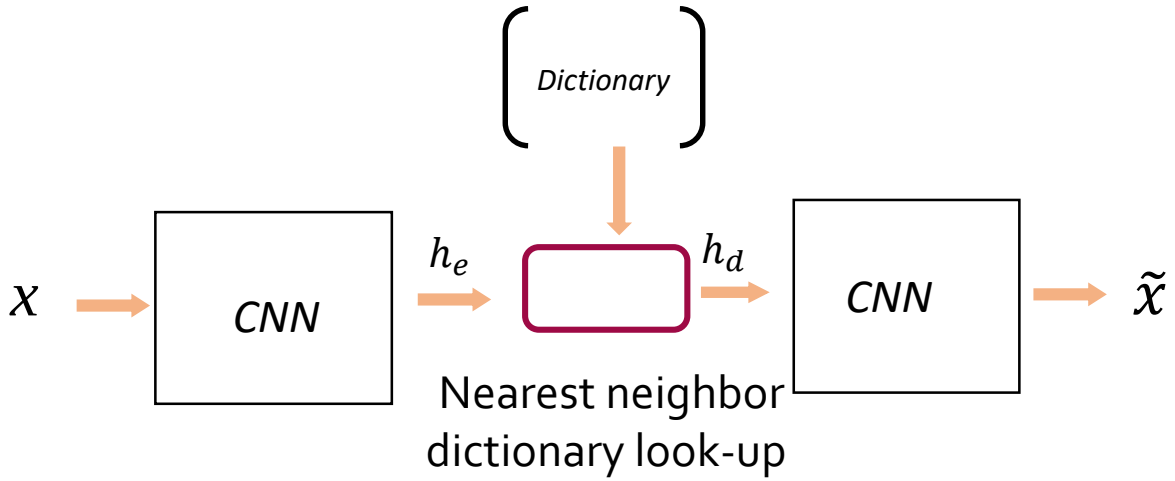
Proposed Sparse Representation Learning

- 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]



[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.

Proposed Sparse Representation Learning

- Training

  - Encoder Training: Inspired by the ISTA algorithm

$$\min_{D} \frac{1}{T} \sum_{i=1}^{T} \min_{h^{(i)}} \frac{1}{2} \left\| x^{(i)} - Dh^{(i)} \right\|_2^2 + \lambda \left\| h^{(i)} \right\|_1 \Rightarrow \begin{cases} (1) \quad h_{t+1}^{temp} = h_t - \alpha \nabla \| x - Dh_t \|_2^2 \\ (2) \quad h_{t+1} = shrink(h_{t+1}^{temp}, \lambda\alpha) \end{cases}$$

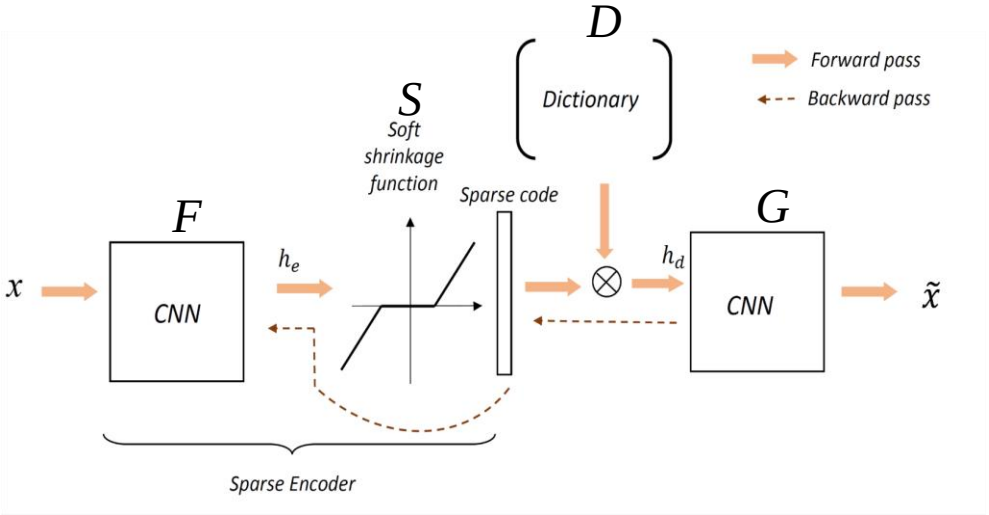
minimize ℓ_2 distance

Apply ℓ_1 norm

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

Backward pass

Forward pass



Experiments and Results

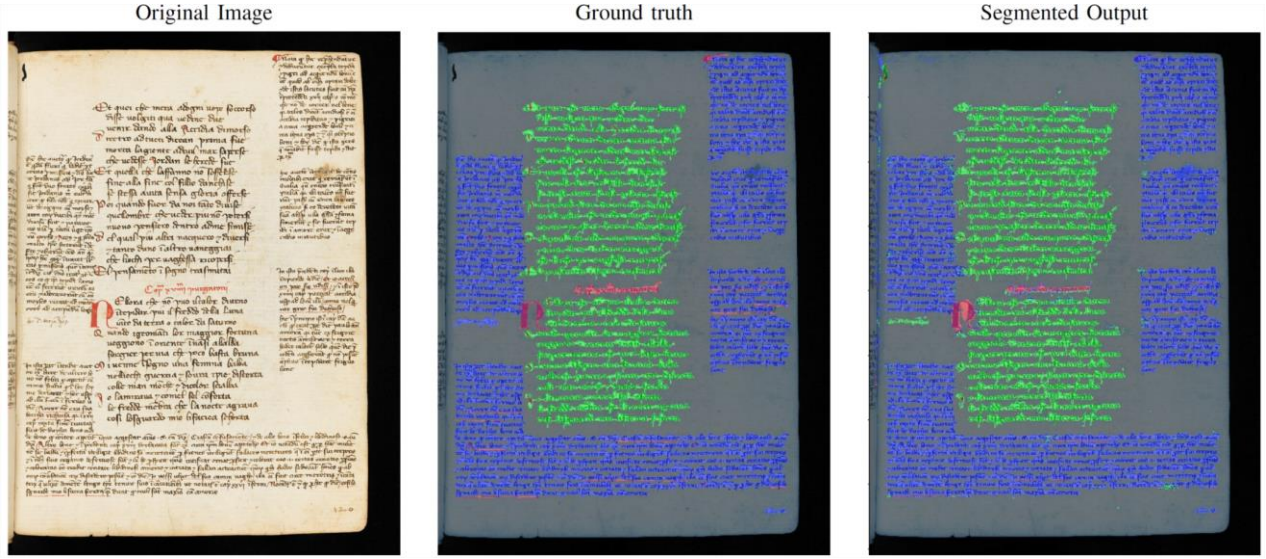
- DIVA-HisDB dataset

○ 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

	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	_	_

Thanks for Your Attention

Waiting to meet you virtually and discuss more!