Multimodal Extension of the ML-CSC Framework for Medical Image Segmentation

Jens Janssens[†], Srđan Lazendić[*,†], Shaoguang Huang[†] and Aleksandra Pižurica[†]
[†]Department of Telecommunications and Information Processing, GAIM,
[*]Department of Electronics and Information Systems, Clifford Research Group,
Faculty of Engineering and Architecture, Ghent University, Belgium
{Jens.Janssens; Srdan.Lazendic; Shaoguang.Huang; Aleksandra.Pizurica}@UGent.be

***Abstract*—In recent years, Convolutional Neural Networks (CNNs) have led to huge successes across various computer vision applications. However, the lack of interpretability poses a severe barrier for their wider adoption in healthcare. Recently introduced Multilayer Convolutional Sparse Coding (ML-CSC) data model provides a model-based explanation of CNNs. This article aims to extend the ML-CSC framework towards multimodal data processing, which to our knowledge has not been addressed so far. In particular, we focus on interpretable medical image segmentation architecture design for multimodal data. We derive a novel sparse coding algorithm and propose three different CNN architectures with increasing performance, without introducing any additional learnable parameters. Based on the sparse coding theory, our multimodal extension enables the systematic design of interpretable CNN segmentation architectures. Experimental analysis demonstrates that the achieved segmentation results are consistent with the obtained theoretical expectations.**

***Index Terms*—Multilayer convolutional sparse coding, interpretable convolutional neural networks, multimodal data, medical image segmentation**

I. Introduction

Deep learning models have led to many successes in the past decade in different fields of science and engineering, ranging from computer vision to natural language processing [1]. However, for a particular type of problems, designing neural network architectures is often driven by a considerable amount of intuition of deep learning experts and by trial-and-error strategies [2].

This lack of transparency and interpretability poses severe ethical and legal implications when adopting Artificial Intelligence in healthcare [3]. Accordingly, the development of methods for interpretable deep learning models has recently attracted increasing attention. The sparse representation model has led to unprecedented results in the previous years for a wide variety of applications in image and video restoration, content analysis and many others [4]. Deep learning, as an instance of general representation learning, is naturally connected to sparse representations. Recent advances based on a multilayer convolutional extension of the classical sparse representation model give theoretical insights into the success of deep learning models and in particular CNNs. A multilayer extension of the Convolutional Sparse Coding, also known as the Multilayer Convolutional Sparse Coding (ML-CSC), has raised insightful connections between sparse representations and Convolutional Neural Networks (CNNs) [5]. It leads to a solid and systematic theoretical justification of the architectures used in deep learning for CNNs, allowing theoretical analysis of existing CNN architectures and the generation of new ones in a more systematic fashion. The designed architectures are by construction interpretable and more transparent as the design process solely relies on sparse coding theory and it is not driven anymore by intuition or empirical validation. A major constraint of these interpretable CNNs compared to state-of-the-art CNN architectures is that they do not rely on any advanced deep learning regularization techniques, such as pooling and batch normalisation. Hence, they are not yet able to compete with state-of-the-art results. Instead, they focus on enhancing the interpretability of CNN models.

Since contemporary CNNs lack interpretability and transparency, medical experts are hesitant about adopting such tools in their daily workflow. It is well-known that medical imaging tasks benefit from various imaging modalities comprising different and complementary information [6]. Given the need for interpretable CNNs in the medical community, in this article we extend the ML-CSC model towards multimodal data and apply it to the problem of medical image segmentation. The proposed data model provides features at multiple abstraction levels for each segmentation class. To recover these features, we provide multimodal extension of an existing image decomposition sparse coding algorithm. Moreover, by relying on the connections between sparse representations and CNNs, the derived sparse coding algorithms allow us to systematically design CNN feature extraction architectures. The experimental results will be performed on the Brain Tumor Segmentation (BraTS) [7]–[9] dataset to analyse the segmentation performance and verify the consistency with theoretical expectations.

The organisation of the paper is as follows. In Section II we introduce the preliminaries required to build and scaffold the theory in this paper. Section III presents the problem formulation and our multimodal extension of the ML-CSC framework for medical image segmentation. In Section IV we discuss the experimental results on the BraTS dataset. Section V concludes the paper.

II. Preliminaries

A. Multilayer Convolutional Sparse Model

The sparse representation model for an image patch $\mathbf{x} \in \mathbb{R}^n$ is formally defined as $\mathbf{x} = \mathbf{D}\gamma$, where $\gamma \in \mathbb{R}^m$ denotes the sparse representation w.r.t. to an overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$ [4]. The task of recovering γ from $\mathbf{x}$ is better known as sparse coding. The global extension of the classical sparse coding model has been recently introduced under the name of Convolutional Sparse Coding (CSC) [10]. In the CSC model a global signal can be approximated as a linear combination of convolutions between a particular kernel and a convolutional sparse feature map, which can formally be expressed as

$$\mathbf{x} = \sum_{i=1}^{m} \mathbf{K}_i \gamma_i = \mathbf{D}\gamma, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ denotes the vectorised input signal and $\mathbf{K}_i \in \mathbb{R}^{N \times N}$ denotes the Toeplitz or convolutional matrix. Recently, this model was further extended to a multilayer setting where the same structure is recursively imposed on each sparse representation: $\gamma_{i-1} = \mathbf{D}_i \gamma_i$, which leads to the following Multilayer CSC (ML-CSC) model:

$$\mathbf{x} = \mathbf{D}_1 \mathbf{D}_2 \dots \mathbf{D}_L \gamma_L, \tag{2}$$

where γ_L denotes the sparse representation at the deepest layer L and $\mathbf{D}_i$ is the convolutional dictionary at layer i [5]. The main benefit of this multilayer extension is that it enables representation learning at multiple abstraction levels. Note that this is similar to the hierarchy of features learned by a CNN. Moreover, it was shown that using the Soft-Thresholding (ST) algorithm to learn the sparse representation at each layer separately is equivalent to the forward pass in a vanilla CNN [5]. This algorithm, which is known as the *layered ST*, provides the solution of the layer-wise relaxation of the model, i.e. when $\gamma_{i-1} = \mathbf{D}_i \gamma_i$, for $i \in \{1, \dots, L\}$ and $\gamma_0 = \mathbf{x}$. For a two-layer ML-CSC instance, it can be expressed as

$$\hat{\gamma}_2 = \mathcal{S}_{\lambda_2}(\mathbf{D}_2^T \mathcal{S}_{\lambda_1}(\mathbf{D}_1^T \mathbf{x})), \tag{3}$$

where the soft-thresholding operator $\mathcal{S}_\lambda(\cdot)$ with threshold $\lambda \in \mathbb{R}$ is a sparsifying operator, which proves to be equivalent to the ReLU activation function [5].

Given that soft-thresholding only provides a crude approximation, we could alternatively use the theoretical superior ISTA algorithm at each layer, which yields the *layered ISTA* algorithm [5]. However, as both layered ST and ISTA only provide a solution to the layer-wise relaxation of the ML-CSC model, even better results could be obtained by using the *multilayer ISTA* (ML-ISTA) algorithm [11]. This algorithm does not rely on such a layer-wise relaxation since it directly recovers the deepest sparse representation γ_L from the image $\mathbf{x}$ without relying on the (approximately) recovered intermediate representations $\{\hat{\gamma}_i\}_{i=1}^{L-1}$ to compute γ_L.

B. Morphological Component Analysis

In order to separate different features contained in our data, we will rely on Morphological Component Analysis (MCA), which is a frequently used image decomposition method based on sparse representations of unimodal signals [12]. MCA assumes that each signal is a linear mixture of K morphologically distinct components. Formally, the MCA data model for a signal $\mathbf{x}$ is defined as

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{\Phi}_k \gamma_k, \tag{4}$$

where the sparse representation of each component w.r.t. the component dictionary $\mathbf{\Phi}_k$ is denoted as γ_k. The dictionary $\mathbf{\Phi}_k$ has a discriminative role as it allows a sparse representation for each individual component k.

The MCA decomposition algorithm [12], which aims at recovering $\{\gamma_k\}_{k=1}^K$, is an iterative procedure where each iteration solves a coordinate relaxation of the pursuit problem corresponding to Eq. (4). This entails first recovering γ_k while the other coordinates $\{\gamma_{\tilde{k}}\}_{\tilde{k} \neq k}$ are fixed, and this process is repeated for all other coordinates until all sparse representations are computed. To ensure that in early iterations, only the most prominent features are being extracted from $\mathbf{x}$, the employed threshold to recover $\{\gamma_k\}_{k=1}^K$ at each decomposition iteration is linearly decreased from an initially large value towards a stopping threshold $\lambda_{\min}$, which is usually set directly proportional to the noise in $\mathbf{x}$. This approach corresponds to a salient-to-fine feature learning process, where subsequent iterations add progressively more details to the components.

C. Multimodal Convolutional Sparse Coding

A multimodal CSC data model was proposed by Song et al. [13] and applied to image super-resolution. They model the two available modalities $\mathbf{x}$ and $\mathbf{y}$ as a sum of two signals corresponding to the common features and unique features, respectively:

$$\begin{aligned} \mathbf{x} &= \mathbf{\Psi}_c \mathbf{z} + \mathbf{\Psi}\mathbf{u}, \\ \mathbf{y} &= \mathbf{\Phi}_c \mathbf{z} + \mathbf{\Phi}\mathbf{v}. \end{aligned} \tag{5}$$

Vector $\mathbf{z}$ is the joint sparse representation corresponding to the shared features among the two modalities, while the vectors $\mathbf{u}$ and $\mathbf{v}$ correspond to the sparse representations of the unique features of modality $\mathbf{x}$ and $\mathbf{y}$, respectively. The convolutional dictionaries $\mathbf{\Psi}_c$ and $\mathbf{\Phi}_c$ are associated with the joint sparse representation $\mathbf{z}$, whereas convolutional dictionaries $\mathbf{\Psi}$ and $\mathbf{\Phi}$ correspond to the sparse representations $\mathbf{u}$ and $\mathbf{v}$, respectively. However, this model did not yet exploit the power of having representations at multiple abstraction levels as in ML-CSC.

D. Brain Tumor Segmentation Dataset

One of the most widely adopted datasets in recent years for multimodal medical image segmentation is the neuro-oncological Brain Tumor Segmentation (BraTS) dataset [7]–[9]. The BraTS dataset consists of four brain MRI modalities (T1, T1ce, T2 and FLAIR) acquired using different MRI-imaging methods. All modalities are co-registered,

skull-stripped and have the same dimensions. The goal of BraTS is to perform semantic image segmentation, a.k.a. pixel-wise classification. To each pixel, one of the four following classes has to be assigned: Healthy Brain Tissue (HBT), Whole Tumor (WT), Tumor Core (TC) or Enhancing Tumor (ET).

III. Proposed Methods

To demystify the CNN design process for multimodal medical image segmentation and to obtain interpretable CNNs, the ML-CSC model will first be extended towards multimodal data. In particular, the possible extension should aim at modelling the dependencies between the modalities $\{\mathbf{x}^{(i)}\}_{i=1}^{4}$ of the BraTS dataset, and it should facilitate semantic image segmentation for four classes. For the multimodal ML-CSC extension, appropriate pursuit problems and sparse coding algorithms will be derived to recover the sparse representations at multiple abstraction levels. The obtained sparse coding algorithms can then be used for a systematic design of CNN feature extraction architectures. Finally, in order to to perform the pixel-wise prediction and generate the segmentation masks, we will add a classification layer on top of the feature extraction CNN.

A. Data Model

To provide a multimodal extension of the interpretable ML-CSC framework presented in Section II-A, a sparse model for the multimodal data will be first defined. Based on MCA decomposition and the multimodal CSC application presented in Section II-B and II-C, respectively, we start from the following two model assumptions to model the dependencies among the modalities and facilitate segmentation:

- We assume that each modality can be modelled as a linear mixture of four morphological components, where each component corresponds to segmentation class-specific features.
- Since all modalities capture the same underlying phenomenon, they are homogeneous and co-registered, we will assume that the hidden sparse representations of each component are shared among all modalities. In other words, we will assume that the features of each segmentation class are shared between all modalities.

These two assumptions are formalised as follows:

$$\mathbf{x}^{(i)} = \sum_{c=1}^{4} \mathbf{\Phi}_c^{(i)} \gamma_c, \quad i \in \{1,2,3,4\}, \tag{6}$$

where the convolutional dictionary for the i-th modality and component c is given by $\mathbf{\Phi}_c^{(i)}$. The joint sparse representations w.r.t. the convolutional dictionaries for component c, or equivalently, segmentation class c, are denoted by γ_c. Note that the dictionaries $\mathbf{\Phi}_c^{(i)}$ are modality-dependent, enabling us to capture the pixel value differences in the modalities due to the modality-specific MRI-imaging methods. We visualise the data model in Fig. 1.

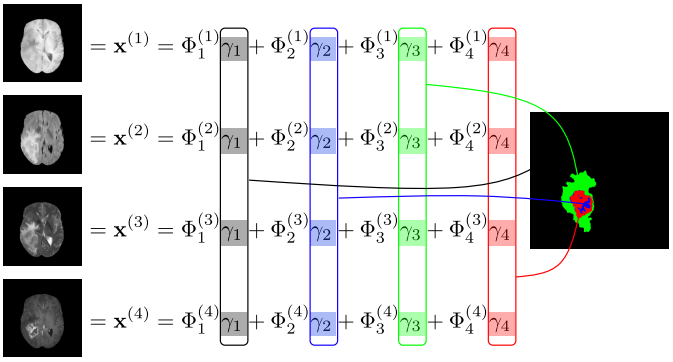


Fig. 1: An illustration of the multimodal BraTS data with its constituting morphological components.

To obtain joint sparse representations of the data at multiple abstraction levels we employ four ML-CSC instances to model each joint sparse representation as

$$\gamma_c = \mathbf{D}_1^{(c)}\mathbf{D}_2^{(c)}\dots\mathbf{D}_l^{(c)}\dots\mathbf{D}_L^{(c)}\gamma_{c,L}, \tag{7}$$

where $\mathbf{D}_l^{(c)}$ denotes the convolutional dictionary for segmentation class c at layer l and the sparse representation at the deepest layer L, or highest level of abstraction, is denoted by $\gamma_{c,L}$. By combining Equations (6) and (7), we obtain a multimodal extension of the classical ML-CSC model (2).

Our proposed data model consists of two modules. The Joint Feature Extraction Module (JFEM) corresponds to the decomposition step in Eq. (6). This module aims to extract, from all modalities $\mathbf{x}^{(i)}$, a joint sparse representation γ_c for each segmentation class c. Next, a higher-level representation can be obtained for these joint representations by an L-layer ML-CSC instance that we refer to as the ML-CSC module. These modules are depicted in Fig. 2.

B. Joint Feature Extraction Module Architecture

1) Pursuit Problem: The task to recover the joint sparse representations $\{\gamma_c\}_{c=1}^{4}$ can formally be expressed as the following optimisation problem in its Lagrangian form:

$$\min_{\{\gamma_c\}_{c=1}^{4}} \underbrace{\frac{1}{2}\sum_{i=1}^{4}\left\|\mathbf{x}^{(i)} - \sum_{c=1}^{4}\mathbf{\Phi}_c^{(i)}\gamma_c\right\|_2^2}_{\text{modalities reconstruction penalty}} + \underbrace{\sum_{c=1}^{4}\lambda_c\|\gamma_c\|_{0,\infty}}_{\text{sparsity constraint}}, \tag{8}$$

where the first term represents a sum of ℓ_2-norm based reconstruction penalties for each modality. As we are operating in a convolutional setting, the local $\ell_{0,\infty}$-pseudonorm, is required to enforce sparsity of the hidden joint representations γ_c [10], with λ_c being the Lagrangian multipliers. As the $\ell_{0,\infty}$-pseudonorm in Eq. (8) makes the optimisation problem NP-hard, a convex relaxation of the non-convex $\ell_{0,\infty}$-pseudonorm towards the convex ℓ_1-norm has to be performed and thus we obtain:

$$\min_{\{\gamma_c\}_{c=1}^{4}} \frac{1}{2}\sum_{i=1}^{4}\left\|\mathbf{x}^{(i)} - \sum_{c=1}^{4}\mathbf{\Phi}_c^{(i)}\gamma_c\right\|_2^2 + \sum_{c=1}^{4}\lambda_c\|\gamma_c\|_1. \tag{9}$$

TABLE I: Relationship between the elementary units in sparse coding algorithms and CNNs.

Sparse coding			CNN
transposed convolutional dictionary operator	$\mathbf{D}^T(\cdot)$	$\leftrightarrow$	convolutional layer
convolutional dictionary operator	$\mathbf{D}(\cdot)$	$\leftrightarrow$	transposed convolutional layer
soft thresholding operator with threshold λ	$\mathcal{S}_\lambda(\mathbf{z})$	$\leftrightarrow$	$\mathrm{ReLU}(\mathbf{z}-\lambda)$

Motivated by the connections with MCA decomposition [12], we also perform a coordinate relaxation of the pursuit problem (9). To simplify the notation, we introduce the marginal residual modalities w.r.t. segmentation class c as $\hat{\mathbf{x}}_c^{(i)} = \mathbf{x}^{(i)} - \sum_{\tilde{c}\neq c} \mathbf{\Phi}_{\tilde{c}}^{(i)} \gamma_{\tilde{c}}$. The coordinate-relaxed pursuit problem can then formally be expressed for every $c \in \{1,2,3,4\}$ as:

$$\min_{\gamma_c} \frac{1}{2} \sum_{i=1}^{4} \left\| \hat{\mathbf{x}}_c^{(i)} - \mathbf{\Phi}_c^{(i)} \gamma_c \right\|_2^2 + \lambda_c \|\gamma_c\|_1. \tag{10}$$

2) Sparse Coding Algorithm: The pursuit problem in Eq. (10) can be solved by a proximal gradient method, which is an iterative procedure consisting of a gradient step succeeded by a proximal mapping [14]. This approach leads to the following update rule aiming at iteratively finding the solution for the pursuit problem in Eq. (10):

$$\gamma_c^{k+1} = \mathcal{S}_{\mu_c \lambda_c} \left[\gamma_c^k - \mu_c \sum_{i=1}^{4} \mathbf{\Phi}_c^{(i)^T} \left(\mathbf{\Phi}_c^{(i)} \gamma_c^k - \hat{\mathbf{x}}_c^{(i)} \right) \right], \tag{11}$$

where μ_c denotes the step size. This update rule serves as our proposed sparse coding algorithm to obtain refined estimates of the higher-level joint sparse representation γ_c from the marginal residual modalities $\hat{\mathbf{x}}_c^{(i)}$ w.r.t. segmentation class c. Note that when removing the sum operator in Equation (11), we obtain the classical ISTA algorithm. Therefore, the derived sparse coding algorithm can be interpreted as a multimodal extension of ISTA, and it will be called MM-ISTA.

However, as the derived MM-ISTA algorithm only computes the minimisers of the *coordinate-relaxed* pursuit problem in Eq. (10), we still need to provide a solution to the overall pursuit problem in Eq. (9) in order to recover *all* higher-level joint sparse representations $\{\gamma_c\}_{c=1}^4$. To this end, a salient-to-fine feature extraction procedure can be adopted to progressively learn the morphological components from the modalities. More detailed features are progressively added to γ_c by linearly decreasing the employed thresholds in subsequent iterations towards a stopping value $\lambda_{\min,c}$. A multimodal extension of the threshold initialisation scheme has to be provided, but due to space limitation, details of this algorithm are omitted here and will follow in the future work.

3) CNN Architecture: Our proposed algorithms can now be used to design a CNN architecture for the JFEM in a systematic fashion. The design process comprises two essential steps. First, the elementary sparse coding units in the sparse coding algorithm should be replaced by their CNN counterparts, as summarised in Table I [5], [11]. Next, the sparse coding algorithm should be unfolded for a fixed number of iterations. This entails fixing the number of joint feature extraction iterations and the number of iterations in MM-ISTA for the JFEM module.

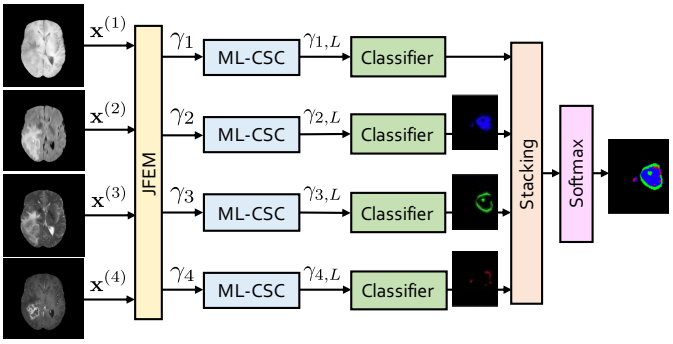


Fig. 2: The proposed segmentation model.

The unfolding technique implies that the same CNN blocks which contain a fixed number of learnable parameters are iterated over time. This replication process allows designing remarkably deep CNNs without an exploding number of parameters, as the learnable parameters are being shared between each iterated block. The purpose of the iterated blocks is to obtain iterative refinements of the features from the data. Further on, the CNN blocks are composed of various (de)convolutional layers, skip connections and summations of feature maps.

C. ML-CSC Module Architecture

1) Pursuit Problem: A missing part in our entire CNN segmentation model is the architecture of the ML-CSC modules (Fig. 2). The purpose of this module is to obtain a higher-level sparse representation $\gamma_{c,L}$ for each joint sparse representation γ_c. The task to recover the higher-level joint sparse representations $\{\gamma_{c,L}\}_{c=1}^4$ can formally be expressed as the following optimisation problem in its Lagrangian form [5], [11]:

$$\begin{aligned} &\min_{\gamma_{c,L}} \frac{1}{2} \left\| \gamma_c - \mathbf{D}_1^{(c)} \mathbf{D}_2^{(c)} \dots \mathbf{D}_L^{(c)} \gamma_{c,L} \right\|_2^2 + \\ &\sum_{l=2}^{L} \lambda_{c,l-1} \left\| \mathbf{D}_l^{(c)} \dots \mathbf{D}_L^{(c)} \gamma_{c,L} \right\|_1 + \lambda_{c,L} \|\gamma_{c,L}\|_1, \end{aligned} \tag{12}$$

where the first term represents an ℓ_2-norm based reconstruction penalty, the second term enforces the sparsity of the intermediate representations while the last term denotes the sparsity constraint for the deepest representation. The Lagrangian multiplier for the l-th layer is denoted by $\lambda_{c,l}$. By introducing the intermediate sparse representations $\gamma_{c,l-1} = \mathbf{D}_l^{(c)} \gamma_{c,l}$ and relaxing the joint ℓ_2-based reconstruction penalty for all L layers, the problem (12) can be simplified into L consecutive CSC problems, which for every $l \in \{1,\dots,L\}$ can be expressed as

$$\min_{\gamma_{c,l}} \frac{1}{2} \left\| \gamma_{c,l-1} - \mathbf{D}_l^{(c)} \gamma_{c,l} \right\|_2^2 + \lambda_{c,l} \|\gamma_{c,l}\|_1, \tag{13}$$

TABLE II: The three considered CNN segmentation models together with the per-class test DSC. The last column reports the average over the three tumor segmentation classes.

Model	JFEM	ML-CSC modules	ET (blue)	WT (green)	TC (red)	Average
JF-L-ST	JFE algorithm	layered ST	0.402 ± 0.282	0.640 ± 0.271	0.603 ± 0.334	0.556 ± 0.312
JF-L-ISTA	JFE algorithm	layered ISTA	0.424 ± 0.305	0.643 ± 0.288	0.613 ± 0.328	0.568 ± 0.319
JF-ML-ISTA	JFE algorithm	ML-ISTA	$\mathbf{0.449 \pm 0.315}$	$\mathbf{0.648 \pm 0.298}$	$\mathbf{0.656 \pm 0.304}$	$\mathbf{0.589 \pm 0.319}$

where $\gamma_{c,0} = \gamma_c$. In this way we obtain problems which can be solved by existing ML-CSC algorithms, which will be briefly discussed.

2) Sparse Coding Algorithm: Optimisation problems of the form (12) and (13) are well-studied in literature. For the layer-wise relaxed pursuit problem (13) the layered Soft Thresholding (ST) and layered ISTA can be employed. A global sparse coding algorithm, the multilayer ISTA (ML-ISTA), provides a solution to the pursuit problem (12) by recovering the sparse representations at each layer all at once instead of the layer-by-layer approach. It is known that the ML-ISTA is theoretically superior w.r.t. representation recovery performance compared to the layered ST and layered ISTA [11]. These three ML-CSC algorithms can now be used to design in a systematic and interpretable way the CNN architecture for the ML-CSC modules.

3) CNN Architecture: Each architecture arises from one of the three discussed ML-CSC algorithms. The design process comprises again two steps. The elementary sparse coding units in the three ML-CSC algorithms should be replaced by their CNN counterparts (Table I). Layered Soft Thresholding CNN leads to the most famous and straightforward CNN architecture, more specifically the LeNet architecture [15]. In the second step, the layered ISTA and ML-ISTA algorithm should be unfolded for a predefined fixed number of iterations. Such an unfolding is not required for layered ST because soft thresholding is not an iterative sparse coding algorithm.

Just as we had in the case of the CNN architecture for JFEM, the unfolding technique implies that the same CNN blocks are iterated over time, for which the parameters in each iterated block are shared. Hence, the interpretable ML-CSC framework is able to construct better performing CNN architectures, by solely relying on sparse coding theory, while keeping the number of parameters constant.

D. Segmentation CNN Architectures

To perform semantic image segmentation, additional processing is required on top of the feature extraction CNN architecture as the main goal is to predict for each pixel an appropriate segmentation class. We based ourselves on the U-Net architecture for semantic image segmentation to perform the pixel-wise prediction [16]. This CNN uses a convolutional layer with a 1×1 kernel as a classification layer to perform the pixel-wise prediction. This could be interpreted as a linear classifier sliding over all pixels in the input feature map $\gamma_{c,L}$ to predict a score for class c. Fig. 2 shows the complete proposed segmentation model. Note that in the end, the predicted binary segmentation masks are stacked, and the softmax activation function is applied.

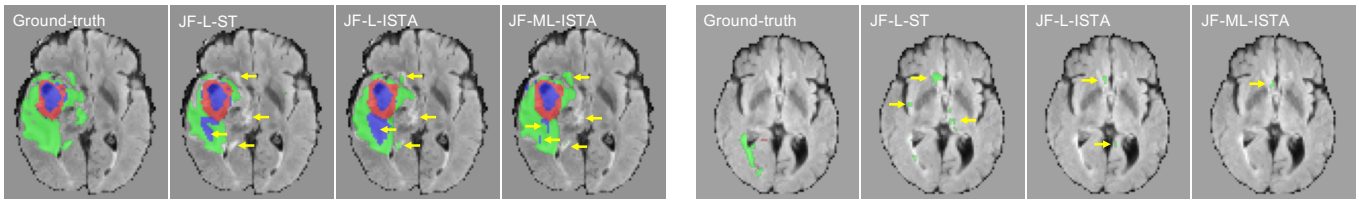
Based on this segmentation model, we propose three increasingly performant CNN models. Table II summarises the sparse coding algorithms used to generate the JFEM and ML-CSC architectures for these models. The CNN architectures for the JFEM and ML-CSC modules have been derived separately by relying on two separate pursuit problems for the data models in Eq. (6) and Eq. (7), respectively. For the JFEM, we derived a Joint Feature Extraction (JFE) sparse coding algorithm, which required us to provide a multimodal extension of the ISTA algorithm. Further, we considered the existing and increasingly performant layered ST, layered ISTA and ML-ISTA architectures for the ML-CSC modules. Therefore, we expect monotonically increasing segmentation performance in the following order: Joint Feature Layered ST (JF-L-ST), Joint Feature Layered ISTA (JF-L-ISTA) and Joint Feature ML-ISTA (JF-ML-ISTA).

IV. Experimental Results and Discussion

So far, it has not yet been specified how the sparse coding parameters, such as the convolutional dictionaries, could be obtained. Therefore, they should be categorised into either learnable parameters or hyperparameters. Motivated by the available literature employing the ML-CSC framework to design interpretable CNNs, we propose to consider the kernels of the convolutional dictionaries $\mathbf{\Phi}_c^{(i)}$ and $\mathbf{D}_l^{(c)}$, the thresholds $\lambda_{\min,c}$ and $\lambda_{c,l}$, and the step sizes μ_c and $\mu_{c,l}$ as learnable parameters [11]. These parameters can be obtained by supervised end-to-end training of the CNNs. All other parameters are deemed hyperparameters, e.g. the number of ML-CSC layers L which equals three in our experiments. Due to the absence of pooling in the ML-CSC framework and the depth of the designed CNNs, we limit the evaluation of our methods to the 75-th axial 2D-slice of the BraTS dataset to keep the training time within reasonable bounds in order to investigate multiple models.

A. Quantitative Test Results

To obtain an unbiased estimate of the segmentation performance, we applied the trained models on the BraTS test dataset. In order to evaluate the performance of the proposed methods per class, we use the Dice Score Coefficient (DSC) evaluation metric. To obtain a general metric over all classes, the average over the per-class DSC is reported. The mean and standard deviation of the DSC for the test samples is reported in Table II. The last column discloses the average over the three tumor segmentation classes.



(a) Example of a successfully segmented sample.

(b) Example of a failed segmentation.

Fig. 3: Qualitative test results for the three considered models together with the ground-truth segmentation mask.

We observe that the obtained segmentation results are consistent with our expectations based on the theoretical results. The DSC increases monotonically when transitioning from the JF-L-ST CNN towards the JF-ML-ISTA CNN for all tumor classes. The most significant performance gain is obtained for the TC class. We can conclude that we were able to improve the segmentation performance by providing systematic improvements to the baseline JF-L-ST CNN model, solely based on sparse coding theory. No extra learnable parameters were introduced in the CNNs. The performance gain has to be attributed to the placement of skip connections and ordering of (de)convolutional layers, theoretically justified by the corresponding sparse coding algorithms.

B. Qualitative Test Results

The segmentation results for two representative BraTS samples are shown in Fig. 3. In Fig. 3(a), the JF-ML-ISTA model clearly performs the best. The leakage of the ET (blue) part into the WT (green) part is substantially smaller for JF-ML-ISTA. For the WT class, the isolated parts are largely missing in all models. Only JF-ML-ISTA succeeds in predicting the isolated part at the top right.

However, for certain samples when the tumor tissue is relatively small in size, the proposed models fail to localise the tumor tissue as shown in Fig. 3(b). Currently, our models cannot exploit the correlations between the adjacent axial slices as we are using 2D convolutions on 2D axial MRI slices. This limitation definitely deserves more attention and will be the scope of future research.

V. Conclusion

In this article, we proposed a multimodal extension of the interpretable ML-CSC framework for medical image segmentation, with the application to the multimodal BraTS dataset. We derived a multimodal ML-CSC model, appropriate pursuit problems, and sparse coding algorithms to find the hidden sparse representations of the data at multiple abstraction levels. The obtained sparse coding algorithms enabled us to systematically design three different CNN segmentation models, with the focus on the interpretability of the designed CNN architectures. Without introducing any additional parameters we were able to increase the segmentation performance. Experiments conducted on the BraTS dataset demonstrate that the obtained segmentation results are consistent with the theoretical expectations.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539 .

[2] M. Elad, D. Simon, and A. Aberdam, "Another step toward demystifying deep neural networks," *Proceedings of the National Academy of Sciences*, vol. 117, no. 44, pp. 27 070–27 072, 2020, doi: 10.1073/pnas.2018957117.

[3] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019, doi: 10.7861/futurehosp.6-2-94.

[4] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010, doi: 10.1007/978-1-4419-7011-4.

[5] V. Papyan, Y. Romano, and M. Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *Journal of Machine Learning Research*, vol. 18, pp. 1–52, 2017, doi: 10.5555/3122009.3176827.

[6] Y. Xu, "Deep learning in multimodal medical image analysis," in *Health Information Science*. Cham: Springer International Publishing, 2019, pp. 193–200.

[7] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, and Y. Burren et al, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015, doi: 10.1109/TMI.2014.2377694.

[8] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, no. 1, p. 170117, 2017, doi: 10.1038/sdata.2017.117.

[9] S. Bakas, M. Reyes, A. Jakab, S. Bauer, and M. R. et al, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," 2019.

[10] V. Papyan, J. Sulam, and M. Elad, "Working locally thinking globally: Theoretical guarantees for convolutional sparse coding," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5687–5701, 2017, doi: 10.1109/TSP.2017.2733447.

[11] J. Sulam, A. Aberdam, A. Beck, and M. Elad, "On Multi-Layer Basis Pursuit, Efficient Algorithms and Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1968–1980, 2020, doi: 10.1109/TPAMI.2019.2904255.

[12] M. Fadili, J.-L. Starck, J. Bobin, and Y. Moudden, "Image decomposition and separation using sparse representations: An overview," *Proceedings of the IEEE*, vol. 98, pp. 983 – 994, 07 2010, doi: 10.1109/JPROC.2009.2024776.

[13] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis, P. L. Dragotti, and M. R. D. Rodrigues, "Multimodal Image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 57–72, 2020, doi: 10.1109/TCI.2019.2916502.

[14] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA, USA: SIAM-Society for Industrial and Applied Mathematics, 2017.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015, doi: 10.1007/978-3-319-24574-4_28.