# On Interpretability of CNNs for Multimodal Medical Image Segmentation

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

***Abstract*—Despite their huge potential, deep learning-based models are still not trustful enough to warrant their adoption in clinical practice. The research on the interpretability and explainability of deep learning is currently attracting huge attention. Multilayer Convolutional Sparse Coding (ML-CSC) data model, provides a model-based explanation of convolutional neural networks (CNNs). In this article, we extend the ML-CSC framework towards multimodal data for medical image segmentation, and propose a merged joint feature extraction ML-CSC model. This work generalizes and improves upon our previous model, by deriving a more elegant approach that merges feature extraction and convolutional sparse coding in a unified framework. A segmentation study on a multimodal magnetic resonance imaging (MRI) dataset confirms the effectiveness of the proposed approach. We also supply an interpretability study regarding the involved model parameters.**

***Index Terms*—Multilayer convolutional sparse coding, interpretable CNNs, multimodal data, medical image segmentation, deep unrolling**

## I. Introduction

Rapid development of deep neural networks has led to many successes in image and data processing [1]. However, tuning the hyperparameters of these sophisticated models typically requires an experienced machine learning expert or practitioner and a considerable amount of intuition and trial-and-error strategies [2].

The lack of transparency and interpretability limits the practical usability of Artificial Intelligence in healthcare [3]. More recently, the explainable deep learning models gained increased attention. Most of the deep learning methods rely on very large training sets and they are considered as black-box models, where the interpretability of the learning process is missing. Techniques called *deep algorithm unrolling* [4], [5] enable a systematic design of deep neural networks based on iterative optimization algorithms. However, many aspects of this approach are yet to be explored, both theoretically and in terms of practical design. Deep learning as an instance of general representation learning is naturally connected to sparse signal representations [6]. Recent advances based on a multilayer convolutional extension of the sparse representation model give theoretical insights into the success of deep learning models [7]. A multilayer extension of the Convolutional Sparse Coding (CSC), also known as the Multilayer CSC (ML-CSC), provides an interpretable framework for studying Convolutional Neural Networks (CNNs), as CNNs can be viewed as an unrolled ML-CSC algorithm [5]. This approach leads to a solid and systematic theoretical justification of the key constituents of CNN models, allowing their theoretical analysis. The designed architectures are by construction interpretable and more transparent as the design process relies on sparse coding theory.

Medical data analysis is one of the domains where explainability and interpretability are particularly important. Hence, despite their huge potential, current deep learning based-models are still not trustful enough to warrant their adoption in clinical practice. The research on explainable models, such as ML-CSC is therefore highly relevant in this domain. In our previous work [8] we have proposed a variant of ML-CSC model for multimodal image segmentation. We are now generalizing that previous approach to an arbitrary number of modalities and segmentation classes. Another important novelty of the present work is a more elegant, unified framework: In [8], the joint feature extraction module (JFEM) and ML-CSC module, were tackled separately. In this work, we will propose a *merged Joint Feature ML-ISTA algorithm* and investigate further in which way the current interpretable ML-CSC framework can be more efficiently extended towards multimodal data. Moreover, we are conducting an interpretability study regarding the different parameters in the proposed CNN segmentation model.

The organization of the paper is as follows. In Section II we introduce preliminaries and related work. The problem formulation of this work is presented in Section III. Section IV presents our joint multimodal extension of the ML-CSC framework. In Section V we discuss the experimental results on the Brain Tumor Segmentation dataset, and we present the results of the conducted interpretability study of the CNNs. Section VI concludes the paper.

## II. Preliminaries

### A. Multilayer Convolutional Sparse Model

For a given image patch represented as a vector $\mathbf{x} \in \mathbb{R}^n$, obtained by stacking the pixel values in a raster scanning fashion and a given dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$ such that $\mathbf{x} = \mathbf{D}\gamma$, the task of recovering the sparse representation $\gamma \in \mathbb{R}^m$ is better known as sparse coding [6]. Recently, an extension of the classical sparse coding model has been introduced by Papyan et al. [9] under the name of Convolutional Sparse

Coding (CSC). The CSC model operates on the entire image $\mathbf{x}$ at once, where the structured dictionaries appearing in the model are given as a concatenation of convolutional dictionaries, formally given by $\mathbf{x} = \sum_{i=1}^{m} \mathbf{d}_i * \gamma_i$, where $\gamma_i$ denotes the convolutional sparse feature maps and $\mathbf{d}_i$ denotes the kernels. Recently, this model was extended to a multilayer setting [7] where the same structure is recursively imposed on each sparse representation: $\gamma_i = \mathbf{D}_{i+1}\gamma_{i+1}$, which leads to the following Multilayer CSC (ML-CSC) model:

$$\mathbf{x} = \mathbf{D}_1\mathbf{D}_2 \dots \mathbf{D}_L\gamma_L, \tag{1}$$

with $\gamma_L$ being the sparse representation at the deepest layer $L$ and $\mathbf{D}_i$ is the convolutional dictionary at layer $i$ [7]. The main benefit of this multilayer extension is that it enables representation learning at multiple abstraction levels which 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 [7]. This algorithm is called *layered ST*. Given that soft-thresholding is a simple algorithm which only provides a crude approximation, in this article we will use use the theoretically and experimentally superior *ML-ISTA* algorithm [10], which does not rely on such a layer-wise relaxation.

### B. Morphological Component Analysis

Morphological Component Analysis (MCA) is a frequently used image decomposition method based on sparse representations of signals [11]. Similar to the sparse representation, the MCA model assumes that each signal is a linear combination of $K$ morphologically distinct components. Formally, the MCA data model can be defined as $\mathbf{x} = \sum_{k=1}^{K} \mathbf{\Phi}_k\gamma_k$, where the sparse representation of each component w.r.t. the component dictionary $\mathbf{\Phi}_k$ is denoted as $\gamma_k$. The dictionary $\mathbf{\Phi}_k$ has a discriminative role as it allows a sparse representation for each individual morphological component $k$.

In order to recover sparse coefficients $\{\gamma_k\}_{k=1}^{K}$, the MCA decomposition algorithm can be used. Each sparse vector, a.k.a. coordinate vector, $\gamma_k$ is obtained separately, while the other coordinates $\{\gamma_{\tilde{k}}\}_{\tilde{k}\neq k}$ are kept fixed. This process is repeated until the sparse representation of each component is computed. To ensure that only the most prominent features are being extracted from $\mathbf{x}$ in the first few iterations, the linearly decreasing threshold strategy is used [11].

A multimodal CSC data model was proposed by Song et al. [12], with application to image super-resolution. In our previous work [8], we have addressed a possible extension of the ML-CSC model towards multimodal data. Here we build further on this work, where the joint feature extraction and ML-CSC modules are tackled jointly, as tackling the pursuit for all layers jointly leads to theoretically superior sparse coding algorithms.

## III. Problem Formulation

A multimodal extension of the ML-CSC model offers an interpetable counterpart for CNN-based segmentation of multimodal images. Here, we aim at modelling the dependencies between the different imaging modalities to facilitate semantic image segmentation. For the multimodal ML-CSC extension, appropriate pursuit problems and sparse coding algorithms will be derived. The obtained sparse coding algorithm will be used to design CNN feature extraction and segmentation architectures as the two main parts of our model.

## IV. Proposed Methods

### A. Data Model

First, we will define a sparse model for the multimodal data. Based on MCA model presented in Section II-B and the multimodal CSC [12], we depart from the following assumptions. First, we assume that each modality can be modelled as a linear combination of different morphological components, where each component corresponds to segmentation class-specific features. Second, 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. These two assumptions can be formalised as follows:

$$\mathbf{x}^{(i)} = \sum_{c=1}^{m} \mathbf{\Phi}_c^{(i)}\gamma_c, \quad i \in \{1, \dots, n\}, \tag{2}$$

where the convolutional dictionary for the $i$-th modality and component $c$ is given by $\mathbf{\Phi}_c^{(i)}$, $m$ is the number of different segmentation classes and $n$ is the number of different modalities. Note that compared to our model in [8], here we allow an arbitrary number of imaging modalities and segmentation classes. The joint sparse representations w.r.t. the convolutional dictionaries for component $c$ are denoted by $\gamma_c$. As we are basing ourselves on the MCA model, the dictionaries $\mathbf{\Phi}_c^{(i)}$ are again modality-dependent, enabling us to capture the pixel value differences in the different image modalities.

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

$$\gamma_c = \mathbf{D}_1^{(c)}\mathbf{D}_2^{(c)} \dots \mathbf{D}_l^{(c)} \dots \mathbf{D}_L^{(c)}\gamma_{c,L}, \tag{3}$$

where $\mathbf{D}_l^{(c)}$ denotes the convolutional dictionary for segmentation class $c$ at layer $l$ and the sparse representation at the deepest layer $L$ is denoted as $\gamma_{c,L}$. By merging Eq. (2) and (3), we obtain our proposed sparse model of the multimodal data:

$$\mathbf{x}^{(i)} = \sum_{c=1}^{m} \mathbf{\Phi}_c^{(i)}\mathbf{D}_1^{(c)}\mathbf{D}_2^{(c)} \dots \mathbf{D}_l^{(c)} \dots \mathbf{D}_L^{(c)}\gamma_{c,L}, \tag{4}$$

for every $i \in \{1, \dots, n\}$. Note that Eq. (4) could be interpreted as a hybridisation of MCA and ML-CSC.

In our previous work [8], the joint feature extraction module (JFEM), which corresponds with the extraction step in Eq. (3), and ML-CSC module, which serves to obtain a higher-level representations jointly, were tackled separately, by the proposed JFE algorithm and ML-ISTA, respectively. In this work, we will propose a *merged Joint Feature ML-ISTA algorithm*, motivated by the fact that tackling the pursuit for

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 $\lambda$ | $\mathcal{S}_\lambda(\mathbf{z})$ | $\leftrightarrow$ | $\mathrm{ReLU}(\mathbf{z}-\lambda)$ |

all layers jointly, instead of sequentially, leads to theoretically superior sparse coding algorithms. This algorithm will enable us to design the merged CNN architecture for the JFEM and ML-CSC module.

### B. Feature Extraction CNN Architecture

*1) Pursuit Problem:* For the sake of clarity, let us consider only a two-layer instance, which corresponds with $L = 1$ in Eq. (4). The task to recover the higher-level joint sparse representations $\{\gamma_{c,L}\}_{c=1}^m$ from the modalities $\mathbf{x}^{(i)}$ can be expressed as the following optimisation problem in its Lagrangian form:

$$\min_{\{\gamma_{c,1}\}_{c=1}^m} \frac{1}{2}\sum_{i=1}^{n}\left\|\mathbf{x}^{(i)} - \sum_{c=1}^{m}\mathbf{\Phi}_c^{(i)}\mathbf{D}_1^{(c)}\gamma_{c,1}\right\|_2^2 + \sum_{c=1}^{m}\lambda_c\left\|\mathbf{D}_1^{(c)}\gamma_{c,1}\right\|_1 + \sum_{c=1}^{m}\lambda_{c,1}\|\gamma_{c,1}\|_1, \tag{5}$$

where the first term is a penalty term consisting of a sum of $\ell_2$-norm based reconstruction penalties for each modality, and the second term enforces the sparsity of the intermediate joint sparse representations $\gamma_c = \mathbf{D}_1^{(c)}\gamma_{c,1}$. Further on, the last term in Eq. (5) is the sparsity constraint for the higher-level representations and the Lagrangian multipliers are denoted by $\lambda_c$ and $\lambda_{c,1}$.

For the sake of simplicity, let us introduce the marginal residual modalities w.r.t. segmentation class $c$ as $\hat{\mathbf{x}}_c^{(i)} = \mathbf{x}^{(i)} - \sum_{\tilde{c}\neq c}\mathbf{\Phi}_{\tilde{c}}^{(i)}\mathbf{D}_1^{(\tilde{c})}\gamma_{\tilde{c},1}$. The coordinate-relaxed pursuit problem can then formally be expressed for every $c \in \{1,\ldots,m\}$ as:

$$\min_{\gamma_{c,1}} \frac{1}{2}\sum_{i=1}^{n}\left\|\hat{\mathbf{x}}_c^{(i)} - \mathbf{\Phi}_c^{(i)}\mathbf{D}_1^{(c)}\gamma_{c,1}\right\|_2^2 + \lambda_1\left\|\mathbf{D}_1^{(c)}\gamma_{c,1}\right\|_1 + \lambda_{c,1}\|\gamma_{c,1}\|_1. \tag{6}$$

*2) Sparse Coding Algorithm:* First, we will solve the coordinate relaxed problem in Eq. (6) by using a proximal gradient-mapping method. As the proximal algorithm has a closed form solution, given by a soft-thresholding operator, we obtain the following update rule aiming at iteratively finding the solution for the pursuit problem in Eq. (6):

$$\gamma_{c,1}^{k+1} = \mathcal{S}_{\mu_{c,1}\lambda_{c,1}}\Bigg[\gamma_{c,1}^k - \mu_{c,1}\mathbf{D}_1^{(c)^T}\Bigg[\mathbf{D}_1^{(c)}\gamma_{c,1}^k - \mathcal{S}_{\mu_c\lambda_c}\left(\mathbf{D}_1^{(c)}\gamma_{c,1}^k - \mu_c\sum_{i=1}^{n}\mathbf{\Phi}_c^{(i)^T}\left(\mathbf{\Phi}_c^{(i)}\mathbf{D}_1^{(c)}\gamma_{c,1}^k - \hat{\mathbf{x}}_c^{(i)}\right)\right)\Bigg]\Bigg], \tag{7}$$

where $\mu_{c,1}$ and $\mu_c$ denote step sizes and the thresholds are given by $\lambda_{c,1}$ and $\lambda_c$.

Our proposed algorithm provides, simultaneously, iteratively refined estimates for the joint representation $\gamma_c$ in Eq. (3) and its higher level counterpart $\gamma_{c,L}$. Note that when only a single modality is considered in Eq. (7), we obtain the classical ML-ISTA algorithm. Therefore, the derived sparse coding algorithm will be called MM-ML-ISTA, as it can be interpreted as a multimodal extension of ML-ISTA.

As the derived MM-ML-ISTA algorithm only computes the joint sparse representations for a single class $c$, we still need to provide a solution to the general pursuit problem defined in Eq. (5) in order to recover all higher-level joint sparse representations $\{\gamma_{c,L}\}_{c=1}^m$. Again, we will base ourselves on the MCA decompisition algorithm. We will adopt, a salient-to-fine feature extraction procedure to learn the morphological components from the modalities progressively. More details are progressively added to the learned representations by linearly decreasing the thresholds in consecutive iterations towards a minimal value $\lambda_{\min,c}$. This stopping threshold is typically set directly proportional to the standard deviation of the noise present in the image modalities. We refer to this sparse coding algorithm as merged Joint Feature ML-ISTA (*merged JF-ML-ISTA*) since it provides a solution to Eq. (5), which is the merged pursuit problem of the joint feature extraction model in Eq. (2) and ML-CSC model in Eq. (3). Detailed derivations together with theoretical and convergence guarantees are omitted due to space limitation and will follow in the future work.

*3) CNN Architecture:* Now we will unfold the proposed merged JF-ML-ISTA algorithm to design an interpretable CNN architecture for the analysed problem. This 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. Next, the sparse coding algorithm should be unrolled for a fixed number of iterations. This entails fixing the number of joint feature extraction iterations $T$ and the number of MM-ML-ISTA iterations $M$. By replicating $T-1$ times the joint feature extraction block and by using $M$ unrollings of the MM-ML-ISTA algorithm, the refined estimates of the coordinate vectors for the joint features are obtained. Unrolling technique implies that the same CNN blocks are iterated over time, for which the parameters in each iterated block are shared. Our merged JF-ML-ISTA algorithm theoretically justifies the interconnections and ordering of layers in the parameter-optimized network, offering more interpretability of the learning process.

### C. Segmentation CNN Architectures

As our main goal is to perform semantic image segmentation, i.e., the pixelwise classification, an additional processing step is required as a segmentation class label should be assigned to each pixel to acquire the desired segmentation mask. We will thus consider the U-Net model [13], which represents one of the state-of-the-art models for medical image segmentation. To produce binary segmentation masks for each

TABLE II
TWO 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-ML-ISTA [8] | JFE algorithm | ML-ISTA | 0.449 ± 0.315 | 0.648 ± 0.298 | 0.656 ± 0.304 | 0.589 ± 0.319 |
| Merged JF-ML-ISTA | merged Joint Feature | ML-ISTA algorithm | **0.471 ± 0.326** | **0.667 ± 0.269** | **0.664 ± 0.298** | **0.605 ± 0.309** |

TABLE III
THE LEARNED THRESHOLDS AND STEP SIZES FOR THE MERGED JF-ML-ISTA CNN MODEL.

| Segmentation class | $\lambda_{\min,c}$ | $\mu_c$ | $\lambda_{c,1}$ | $\mu_{c,1}$ | $\lambda_{c,2}$ | $\mu_{c,2}$ | $\lambda_{c,3}$ | $\mu_{c,3}$ |
|---|---|---|---|---|---|---|---|---|
| HBT ($c=1$) | 0.803 | 1.005 | -0.004 | 1.051 | -0.006 | 1.039 | -0.008 | 1.032 |
| ET ($c=2$) | 0.906 | 0.907 | 0.011 | 1.042 | 0.007 | 1.030 | 0.001 | 1.012 |
| WT ($c=3$) | 0.788 | 0.872 | 0.005 | 0.998 | 0.002 | 0.994 | 0.003 | 0.996 |
| TC ($c=4$) | 0.930 | 0.958 | -0.002 | 1.009 | -0.004 | 1.000 | -0.005 | 1.007 |

segmentation class $c$, the obtained sparse representations $\gamma_{c,L}$ at the deepest layer of the ML-CSC module are fed into a convolution layer with a $1 \times 1$ kernel, which serves as a pixel-wise prediction step. In the end, the predicted binary segmentation masks are stacked, and the softmax activation function is applied.

Based on this segmentation model, we compare our proposed merged JF-ML-ISTA CNN model with the JF-ML-ISTA model [8]. To the best of our knowledge, our previous work [8], is the only existing work where the ML-CSC model for image segmentation has been applied to multimodal data and thus will be used for comparison. Table II summarises the sparse coding algorithms used to generate the JFEM and ML-CSC architectures for this model. The merged JF-ML-ISTA algorithm provides a merged CNN architecture for both the JFEM and ML-CSC modules.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

Now we evaluate the proposed approach in a practical scenario. First, we will categorise all the sparse coding parameters into either learnable parameters or hyperparameters. Motivated by the available literature employing the ML-CSC framework to design interpretable CNNs, we 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 $\mu_c$ and $\mu_{c,l}$ as learnable parameters [10]. 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, same as in [8] to obtain a fair comparison. For each model, we regard the same hyperparameters which are tuned for the merged JF-ML-ISTA model and the best results are reported.

### A. Quantitative Test Results

To numerically evaluate the segmentation performance, we applied the trained models on the neuro-oncological Brain Tumor Segmentation (BraTS) dataset [14]. The BraTS dataset consists of four brain magnetic resonance imaging (MRI) modalities (T1, T1ce, T2 and FLAIR) acquired using different MRI-imaging methods. As the objective of BraTS is to perform semantic image segmentation, we need to assign to each pixel, one of the four following classes: Healthy Brain Tissue (HBT), Whole Tumor (WT), Tumor Core (TC), or Enhancing Tumor (ET) and thus we consider $i, c \in \{1, 2, 3, 4\}$. The mean and standard deviation of the Dice Score (DSC) for the 67 test samples is reported in Table II. The average values obtained over the three tumor segmentation classes are presented in the last column. Note that we were able to improve the performance for all tumor classes by providing systematic improvements to the JF-ML-ISTA model, based on sparse coding theory and algorithm unrolling. This was done without introducing any additional parameters in the merged JF-ML-ISTA CNN model.

### B. Interpretability Study of the Model Parameters

The generated CNN architectures are by design, theoretically justified, as they emerge from unrolling an appropriate sparse coding algorithm for a fixed number of iterations. The available literature recommends the training ML-CSC models using the Adam optimiser [10], [15], [16]. However, this approach can lead to inconsistent values w.r.t. sparse coding theory, such as negative thresholds, as has already been reported in literature [8], [15], [16]. Therefore, we conduct an interpretability study to verify whether the theoretical interpretability claims offered by sparse representations are valid in practice.

The learned values for all thresholds and step sizes in the merged JF-ML-ISTA model are shown in Table III. The stopping thresholds and step sizes for the joint feature extraction should stay positive, as implied by the theoretical explanation. Notably, $\lambda_{\min,c} > 0$ and $\mu_c > 0$, as well as the step sizes $\{\mu_{c,l}\}_{l=1}^3$ at the three ML-CSC layers are consistent with the theoretical expectations. However, for the thresholds at the ML-CSC layers $\{\lambda_{c,l}\}_{l=1}^3$, it turns out that the network learned a negative value for the HBT and TC segmentation class.

For the four classifiers used to predict the binary segmentation masks for each class, we opted to employ the same parameters in each classifier. In this way, we are able to force, to a certain extent, the discriminative behaviour of the model to be present in the feature extraction architecture.

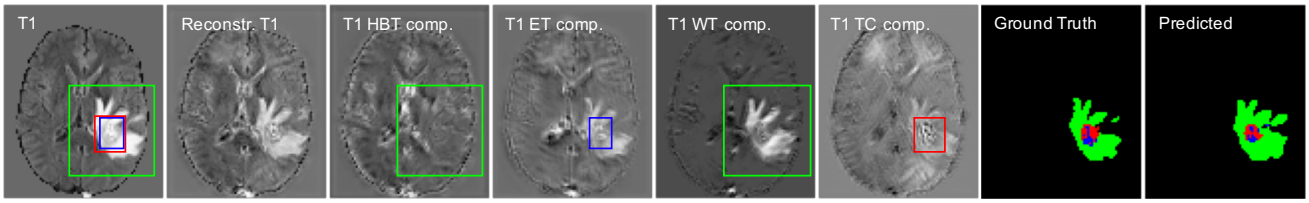


Fig. 1. T1 MRI modality decomposition of the 75th percentile sample. No notable features are visible in HBT component compared to ET, WT and TC class.

The discriminative capabilities of the classifiers decrease by considering similar parameters for each classifier. Hence, in such a manner we are able to produce more discriminative features.

Further on, we also analysed the decomposition of the MRI modalities into four components corresponding to segmentation class-specific features. This constitutes an important step in the representation learning process as it extracts the joint sparse representations $\gamma_c$ for each segmentation class. Fig. 1 shows a successful decomposition for the T1 modality $\mathbf{x}^{(1)}$ of the 75th percentile sample. The four components $\{\mathbf{\Phi}_c^{(1)}\gamma_c\}_{c=1}^4$ are depicted together with the reconstructed modality. We observe that the components were able to extract features from the T1 modality relating to their corresponding segmentation class. Indeed, note that in the HBT component, no noteworthy features remain w.r.t. the ET, WT and TC class. When no notable features relating to the segmentation classes are visible in their corresponding components, the decomposition can lead to unsatisfying results.

We attribute both the inconsistent learned sparse coding parameters and the unsatisfactory decomposition for certain samples to the current end-to-end supervised training procedure. In the current experimental setup, we cannot ensure that the sparse coding parameters in the CNNs converge to values attaining the expected sparse coding functionality, which is vital as, for instance, the learned decomposition dictionaries $\mathbf{\Phi}_c^{(i)}$ constitute a crucial determinant for the decomposition performance [11].

## VI. Conclusion

In this article, we derived a novel multimodal ML-CSC model, appropriate pursuit problems, and sparse coding algorithms. The obtained sparse coding algorithms offer us a systematic way of building the CNN segmentation model, called the Merged Joint Feature Extraction Model. Experimental results conducted on the BraTS dataset demonstrate that we were able to improve the performance for all tumor classes compared to JF-ML-ISTA model, by providing systematic architecture design based on sparse coding and deep unrolling, while keeping the same amount of learnable parameters in both models. Each learnable parameter in our model has a well-defined role in the representation learning process.

## Acknowledgment

This research has been partially supported by the Flanders AI Research Programme grant no. 174B09119.

## 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] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th international conference on international conference on machine learning*, 2010, pp. 399–406.

[5] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.

[6] 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.

[7] 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.

[8] J. Janssens, S. Lazendić, S. Huang, and A. Pižurica, "Multimodal extension of the ML-CSC framework for medical image segmentation," in *12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2021, pp. 91–96.

[9] 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.

[10] 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.

[11] 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.

[12] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis *et al.*, "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.

[13] 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.

[14] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer *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.

[15] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1828–1837.

[16] J. Xiang, Y. Dong, and Y. Yang, "FISTA-Net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1329–1339, 2021.