Towards improving discriminative reconstruction via simultaneous dense and sparse coding

Anonymous Author
Anonymous Institution

Abstract

Discriminative features extracted from the sparse coding model have been shown to perform well for classification and reconstruction. Recent deep learning architectures have further improved reconstruction in inverse problems by considering new dense priors learned from data. We propose a novel dense and sparse coding model that integrates both representation capability and discriminative features. The model considers the problem of recovering a dense vector $\boldsymbol{x}$ and a sparse vector $\boldsymbol{u}$ given measurements of the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$. Our first analysis proposes a natural geometric condition based on the minimal angle between spanning subspaces corresponding to the measurement matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ to establish the uniqueness of solutions to the linear system. The second analysis shows that, under mild assumptions, a convex program recovers the dense and sparse components. We validate the effectiveness of the proposed model on simulated data and propose a dense and sparse autoencoder (DenSaE) tailored to learning the dictionaries from the dense and sparse model. We demonstrate that a) DenSaE denoises natural images better than architectures derived from the sparse coding model ($\boldsymbol{B}\boldsymbol{u}$), b) in the presence of noise, training the biases in the latter amounts to implicitly learning the $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$ model, c) $\boldsymbol{A}$ and $\boldsymbol{B}$ capture low- and high-frequency contents, respectively, and d) compared to the sparse coding model, DenSaE offers a balance between discriminative power and representation.

Preliminary work. Under review by AISTATS 2021. Do not distribute.

1 Introduction

Given a data set, learning a dictionary in which each example admits a sparse representation is tremendously useful in a number of tasks (Aharon et al., 2006; Mairal et al., 2011). This problem, known as sparse coding (Olshausen and Field, 1997) or dictionary learning (Garcia-Cardona and Wohlberg, 2018), has been the subject of significant investigation in recent years in the signal processing community. A growing body of work has mapped the sparse coding problem into encoders for sparse recovery (Gregor and Lecun, 2010), and into autoencoders for purely classification (Rolfe and LeCun, 2013) or denoising (Simon and Elad, 2019; Tolooshams et al., 2020) purposes.

Autoencoders are widely used for unsupervised learning. Their integration with supervised tasks and classifiers has become popular for their regularization power and reduction of the generalization gap (Vincent et al., 2010; Epstein et al., 2018; Epstein and Meir, 2019). Rolfe and LeCun (2013) have shown benefits of autoencoders and sparse features in discriminative tasks.

For data reconstruction, recent work has highlighted some limitations of convolutional sparse coding (CSC) autoencoders (Simon and Elad, 2019) and its multi-layer and deep generalizations (Sulam et al., 2019; Zazo et al., 2019). Simon and Elad (2019) argue that the sparsity levels that CSC allows can only accommodate very sparse vectors, making it unsuitable to capture all features of signals such as natural images, and propose to compute the minimum mean-squared error solution under the CSC model, which is a dense vector capturing a richer set of features.

To address the aforementioned limitations of classical sparse coding, we propose a dense and sparse coding model that represents a signal as the sum of two components: one that admits a dense representation $\boldsymbol{x}$ in a dictionary $\boldsymbol{A}$ that is useful for reconstruction, and another whose representation $\boldsymbol{u}$ is discriminative and sparse in a second dictionary $\boldsymbol{B}$. Based on empirical evidence, the authors in (Zazo et al., 2019) argue that a multi-layer extension of this model can, in principle,

have arbitrary depth. However, to our knowledge, the dense and sparse coding model has not been yet fully analyzed. Our contributions are

Conditions for identifiability and recovery by convex optimization: We derive conditions under which the dense and sparse representation is unique. We then propose a convex program for recovery that minimizes $||\boldsymbol{A}\boldsymbol{x}||_2^2 + ||\boldsymbol{u}||_1$, subject to linear constraints.

Phase-transition curves: We demonstrate through simulations that the convex program can successfully solve the dense and sparse coding problem.

Discriminative reconstruction: We propose a dense and sparse autoencoder (DenSaE) that has competitive discriminative power and improves the representation capability compared to sparse networks.

Organization: **Section 2** discusses the theoretical analysis of the dense and sparse coding problem. Phase transition, classification, and denoising experiments appear in **Section 3**. We conclude in **Section 4**.

1.1 Related work

We comment on the most closely related models. Given the measurements $\boldsymbol{y}$, the problem of recovering $\boldsymbol{x}$ and $\boldsymbol{u}$ is similar in flavor to sparse recovery in the union of dictionaries (Donoho and Huo, 2001; Elad and Bruckstein, 2002; Donoho and Elad, 2003; Soltani and Hegde, 2017; Studer et al., 2011; Studer and Baraniuk, 2014). Most results in this literature take the form of an uncertainty principle that relates the sum of the sparsity of $\boldsymbol{x}$ and $\boldsymbol{u}$ to the mutual coherence between $\boldsymbol{A}$ and $\boldsymbol{B}$, and which guarantees that the representation is unique and identifiable by ℓ_1 minimization. To our knowledge, the analysis of this program is novel and in sharp contrast to classical settings in sparse approximation, in which the objective consists of a single sparsifying norm, rather than the combination of different norms. Robust PCA (Candès et al., 2011), which decomposes a matrix as the sum of low-rank and sparse matrices, uses the combination of the ℓ_1 and nuclear norms, giving it a flavor similar to our problem.

Our model resembles weighted LASSO (Lian et al., 2018; Mansour and Saab, 2017). Compared to weighted LASSO, we can directly map the weighted LASSO objective $\|\boldsymbol{W}\boldsymbol{\alpha}\|_1$ to $\|\boldsymbol{u}\|_1$ by letting $\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{x} & \boldsymbol{u}\end{bmatrix}^T$ and choosing appropriately the entries of a diagonal matrix $\boldsymbol{W}$, with $W_{ij} \in \{0, 1\}$; however, in the weighted LASSO formulation, constraints can only be enforced on the sparse component $\boldsymbol{u}$. Our work differs in that a significant part of our analysis is the directed Euclidean norm constraint on $\boldsymbol{x}$, which recovers a unique solution $\boldsymbol{x}^\star \in \mathrm{Ker}(\boldsymbol{A})^\perp$. Our model can also be interpreted as a special case of Morphological Component Analysis (MCA) (Elad et al., 2005) for $K = 2$, $\boldsymbol{s} = \sum_{k=1}^{K} \boldsymbol{\Phi}_k \boldsymbol{\alpha}_k$, with, however, some distinct differences: i) MCA encodes different morphological structures via the dictionaries $\boldsymbol{\Phi}_k$. We encode a smooth morphological component via the whole product $\boldsymbol{A}\boldsymbol{x}$, which is conceptually different, and ii) we make no assumption of sparsity on the dense component $\boldsymbol{x}$. This leads to an optimization objective that is the combination of ℓ_1 and ℓ_2 norms, unlike that of MCA. Finally, a bare application of noisy sparse coding would treat $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{x}$ as arbitrary noise, hence i) recovers $\boldsymbol{u}$ approximately and ii) cannot recover $\boldsymbol{x}$. However, in our analysis, the term $\boldsymbol{A}\boldsymbol{x}$ is not just undesired noise but represents a sought-out feature. We can recover both $\boldsymbol{x}$ and $\boldsymbol{u}$ exactly. See **Appendix C** for a comparison of our model to noisy compressive sensing. We note that the full dense and sparse coding model is $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{e}$ where $\boldsymbol{e}$ is Gaussian noise.

Notation. Lowercase and uppercase boldface letters denote column vectors and matrices, respectively. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$ and a support set $S \subset \{1, ..., n\}$, $\boldsymbol{x}_S$ denotes the restriction of $\boldsymbol{x}$ to indices in S. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times p}$, $\boldsymbol{A}_S$ is a submatrix of size $m \times |S|$ with column indices in S. The column space of a matrix $\boldsymbol{A}$ (the span of the columns of $\boldsymbol{A}$) is designated by $\mathrm{Col}(\boldsymbol{A})$, its null space by $\mathrm{Ker}(\boldsymbol{A})$. We denote the Euclidean, ℓ_1 and ℓ_∞ norms of a vector, respectively as $||\boldsymbol{x}||_2$, $||\boldsymbol{x}||_1$, and $||\boldsymbol{x}||_\infty$. The operator and infinity norm of a matrix $\boldsymbol{A}$ are respectively denoted as $||\boldsymbol{A}||$ and $||\boldsymbol{A}||_\infty$. The sign function, applied componentwise to a vector $\boldsymbol{x}$, is denoted by $\mathrm{sgn}(\boldsymbol{x})$. The indicator function is denoted by $\mathbb{1}$. The column vector $\boldsymbol{e}_i$ denotes the vector of zeros except a 1 at the i-th location. The orthogonal complement of a subspace $\boldsymbol{W}$ denoted by $\boldsymbol{W}^\perp$. The operator $\mathcal{P}_{\boldsymbol{W}}$ denotes the orthogonal projection operator onto the subspace $\boldsymbol{W}$.

2 Theoretical Analysis

The dense and sparse coding problem studies the solutions of the linear system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$. Given matrices $\boldsymbol{A} \in \mathbb{R}^{m\times p}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ and a vector $\boldsymbol{y} \in \mathbb{R}^m$, the goal is to provide conditions under which there is a unique solution $(\boldsymbol{x}^*, \boldsymbol{u}^*)$, where $\boldsymbol{u}^*$ is s-sparse, and an algorithm for recovering it.

2.1 Uniqueness results for the feasibility problem

In this subsection, we study the uniqueness of solutions to the linear system accounting for the different structures the measurement matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can have. For more details of all the different cases we consider, we refer the reader to **Appendix A**. The main result of this subsection is Theorem 3 which, under a natural geometric condition based on the minimum principal

angle between the column space of $\boldsymbol{A}$ and the span of s columns in $\boldsymbol{B}$, establishes a uniqueness result for the dense and sparse coding problem. Since the vector $\boldsymbol{u}$ in the proposed model is sparse, we consider the classical setting of an overcomplete measurement matrix $\boldsymbol{B}$ with $n \gg m$. The next theorem provides a uniqueness result assuming a certain direct sum representation of the space $\mathbb{R}^m$.

Theorem 1 *Assume that there exists at least one solution to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, namely the pair $(\boldsymbol{x}^*, \boldsymbol{u}^*)$. Let S, with $|S| = s$, denote the support of $\boldsymbol{u}^*$. If $\boldsymbol{B}_S$ has full column rank and $\mathbb{R}^m = \mathrm{Col}(\boldsymbol{A}) \oplus \mathrm{Col}(\boldsymbol{B}_S)$, the only unique solution to the linear system, with the condition that any feasible s-sparse vector $\boldsymbol{u}$ is supported on S and any feasible $\boldsymbol{x}$ is in $\mathrm{Ker}(\boldsymbol{A})^{\perp}$, is $(\boldsymbol{x}^*, \boldsymbol{u}^*)$.*

Proof 1 *Let $(\boldsymbol{x}, \boldsymbol{u})$, with $\boldsymbol{u}$ supported on S and $\boldsymbol{x}^* \in \mathrm{Ker}(\boldsymbol{A})^{\perp}$, be another solution pair. It follows that $\boldsymbol{A}\boldsymbol{\delta_1} + \boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \boldsymbol{0}$ where $\boldsymbol{\delta_1} = \boldsymbol{x} - \boldsymbol{x}^*$ and $\boldsymbol{\delta_2} = \boldsymbol{u}_S - \boldsymbol{u}_S^*$. Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal bases of $\mathrm{Col}(\boldsymbol{A})$ and $\mathrm{Col}(\boldsymbol{B}_S)$ respectively. The equation $\boldsymbol{A}\boldsymbol{\delta_1} + \boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \boldsymbol{0}$ can equivalently be written as $\sum_{i=1}^{r}\langle \boldsymbol{A}\boldsymbol{\delta_1}, \boldsymbol{U}_i\rangle \boldsymbol{U}_i + \sum_{i=1}^{q}\langle \boldsymbol{B}_S(\boldsymbol{\delta_2})_S, \boldsymbol{V}_i\rangle \boldsymbol{V}_i = \boldsymbol{0}$ with $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ denoting the i-th column of $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively. More compactly, we have $\begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}\begin{bmatrix}\{\langle \boldsymbol{A}\boldsymbol{\delta_1}, \boldsymbol{U}_i\rangle\}_{i=1}^{r} \\ \{\langle \boldsymbol{B}_S(\boldsymbol{\delta_2})_S, \boldsymbol{V}_i\rangle\}_{i=1}^{q}\end{bmatrix} = \boldsymbol{0}$. Noting that the matrix $\begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}$ has full column rank, the homogeneous problem admits the trivial solution implying that $\boldsymbol{A}\boldsymbol{\delta_1} = \boldsymbol{0}$ and $\boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \boldsymbol{0}$. Since $\boldsymbol{B}_S$ has full column rank and $\boldsymbol{\delta_1} \in \{\mathrm{Ker}(\boldsymbol{A}) \cap \mathrm{Ker}(\boldsymbol{A})^{\perp}\}$, it folows that $\boldsymbol{\delta_1} = \boldsymbol{\delta_2} = \boldsymbol{0}$. Therefore, $(\boldsymbol{x}^*, \boldsymbol{u}^*)$ is the unique solution.*

The uniqueness result in the above theorem hinges on the representation of the space $\mathbb{R}^m$ as the direct sum of the subspaces $\mathrm{Col}(\boldsymbol{A})$ and $\mathrm{Col}(\boldsymbol{B}_S)$. We use the definition of the minimal principal angle between two subspaces, and its formulation in terms of singular values (Björck and Golub, 1973), to derive an explicit geometric condition for the uniqueness analysis of the linear system in the general case.

Definition 2 *Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal basis of $\mathrm{Col}(\boldsymbol{A})$ and $\mathrm{Col}(\boldsymbol{B})$ respectively. The minimum principal angle between the subspaces $\mathrm{Col}(\boldsymbol{A})$ and $\mathrm{Col}(\boldsymbol{B})$ is defined as follows*

$$\cos(\mu(\boldsymbol{U}, \boldsymbol{V})) = \max_{\boldsymbol{u}\in\mathrm{Col}(\boldsymbol{U}), \boldsymbol{v}\in\mathrm{Col}(\boldsymbol{V})} \frac{\boldsymbol{u}^T\boldsymbol{v}}{||\boldsymbol{u}||_2||\boldsymbol{v}||_2}, \quad (1)$$

The minimum angle $\mu(\boldsymbol{U}, \boldsymbol{V})$ is also equal to the largest singular value of $\boldsymbol{U}^T\boldsymbol{V}$, $\cos(\mu(\boldsymbol{U}, \boldsymbol{V})) = \sigma_1(\boldsymbol{U}^T\boldsymbol{V})$.

Theorem 3 *Assume that there exists at least one solution to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, namely the pair $(\boldsymbol{x}^*, \boldsymbol{u}^*)$. Let S, with $|S| = s$, denote the support of $\boldsymbol{u}^*$. Assume that $\boldsymbol{B}_S$ has full column rank . Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal bases of $\mathrm{Col}(\boldsymbol{A})$ and $\mathrm{Col}(\boldsymbol{B}_S)$ respectively. If $\cos(\mu(\boldsymbol{U}, \boldsymbol{V})) = \sigma_1(\boldsymbol{U}^T\boldsymbol{V}) < 1$, the only unique solution to the linear system, with the condition that any feasible s-sparse vector $\boldsymbol{u}$ is supported on S and any feasible $\boldsymbol{x}$ is in $\mathrm{Ker}(\boldsymbol{A})^{\perp}$, is $(\boldsymbol{x}^*, \boldsymbol{u}^*)$.*

Proof 2 *Consider any candidate solution pair $(\boldsymbol{x}^* + \boldsymbol{\delta_1}, \boldsymbol{u}^* + \boldsymbol{\delta_2})$. We will prove uniqueness by showing that $\boldsymbol{A}\boldsymbol{\delta_1} + \boldsymbol{B}_S(\boldsymbol{\delta_2})_S = 0$ if and only if $\boldsymbol{\delta_1} = \boldsymbol{0}$ and $(\boldsymbol{\delta_2})_S = \boldsymbol{0}$. Using the orthonormal basis set $\boldsymbol{U}$ and $\boldsymbol{V}$, $\boldsymbol{A}\boldsymbol{\delta_1} + \boldsymbol{B}_S\boldsymbol{u}_S$ can be represented as : $\boldsymbol{A}\boldsymbol{\delta_1} + \boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{\delta_1} \\ \boldsymbol{V}^T\boldsymbol{B}_S(\boldsymbol{\delta_2})_S\end{bmatrix}$. For simplicity of notation, let $\boldsymbol{K}$ denote the block matrix: $\boldsymbol{K} = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}$. If we can show that the columns of $\boldsymbol{K}$ are linearly independent, it follows that $\boldsymbol{A}\boldsymbol{\delta_1} + \boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \boldsymbol{0}$ if and only if $\boldsymbol{A}\boldsymbol{\delta_1} = \boldsymbol{0}$ and $\boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \boldsymbol{0}$. We now consider the matrix $\boldsymbol{K}^T\boldsymbol{K}$ which has the following representation*

$$\begin{aligned}\boldsymbol{K}^T\boldsymbol{K} &= \begin{bmatrix}[\boldsymbol{I}]_{r\times r} & [\boldsymbol{U}^T\boldsymbol{V}]_{r\times q} \\ [\boldsymbol{V}^T\boldsymbol{U}]_{q\times r} & [\boldsymbol{I}]_{q\times q}\end{bmatrix} \\ &= \begin{bmatrix}[\boldsymbol{I}]_{r\times r} & [\boldsymbol{0}]_{r\times q} \\ [\boldsymbol{0}]_{q\times r} & [\boldsymbol{I}]_{q\times q}\end{bmatrix} + \begin{bmatrix}[\boldsymbol{0}]_{r\times r} & [\boldsymbol{U}^T\boldsymbol{V}]_{r\times q} \\ [\boldsymbol{V}^T\boldsymbol{U}]_{q\times r} & [\boldsymbol{0}]_{q\times q}\end{bmatrix}.\end{aligned}$$

With the singular value decomposition of $\boldsymbol{U}^T\boldsymbol{V}$ being $\boldsymbol{U}^T\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{R}^T$, the last matrix in the above representation has the following equivalent form $\begin{bmatrix}\boldsymbol{0} & \boldsymbol{U}^T\boldsymbol{V} \\ \boldsymbol{V}^T\boldsymbol{U} & \boldsymbol{0}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}^T$. It now follows that $\begin{bmatrix}\boldsymbol{0} & \boldsymbol{U}^T\boldsymbol{V} \\ \boldsymbol{V}^T\boldsymbol{U} & \boldsymbol{0}\end{bmatrix}$ is similar *to the matrix $\begin{bmatrix}\boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{bmatrix}$. Hence, the nonzero eigenvalues of $\boldsymbol{K}^T\boldsymbol{K}$ are $1 \pm \sigma_i$, $1 \le i \le \min(p, q)$, with σ_i denoting the i-th largest singular value of $\boldsymbol{U}^T\boldsymbol{V}$. Using the assumption $\sigma_1 < 1$ results the bound $\lambda_{\min}\left(\boldsymbol{K}^T\boldsymbol{K}\right) > 0$. It follows that the columns of $\boldsymbol{K}$ are linearly independent, and hence $\boldsymbol{A}\boldsymbol{\delta_1} = \boldsymbol{0}$ and $\boldsymbol{B}_S(\boldsymbol{\delta_2})_S = \boldsymbol{0}$. Since $\boldsymbol{B}_S$ is full column rank and $\boldsymbol{\delta_1} \in \{\mathrm{Ker}(\boldsymbol{A}) \cap \mathrm{Ker}(\boldsymbol{A})^{\perp}\}$, it follows that $\boldsymbol{\delta_1} = \boldsymbol{0}$ and $(\boldsymbol{\delta_2})_S = \boldsymbol{0}$. This concludes the proof.*

A restrictive assumption of the above theorem is that the support of the sought-after s-sparse solution $\boldsymbol{u}^*$ is known. We can remove this assumption by considering $\mathrm{Col}(\boldsymbol{A})$ and $\mathrm{Col}(\boldsymbol{B}_T)$ where T is an arbitrary subset of $\{1, 2, ..., n\}$ with $|T| = s$. More precisely, we state the following corollary whose proof is similar to the proof of Theorem 3.

Corollary 4 *Assume that there exists at least one solution to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, namely the pair $(\boldsymbol{x}^*, \boldsymbol{u}^*)$. Let S, with $|S| = s$, denote the support of $\boldsymbol{u}^*$ and T be an arbitrary subset of $\{1, 2, ..., n\}$ with $|T| \le s$. Assume*

that any $2s$ columns of $\boldsymbol{B}$ are linearly independent. Let $\boldsymbol{U} \in \mathbb{R}^{m\times p}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal bases of $\text{Col}(\boldsymbol{A})$ *and* $\text{Col}(\boldsymbol{B}_{S\cup T})$ *respectively. If $\mu(\boldsymbol{U},\boldsymbol{V}) = \sigma_1(\boldsymbol{U}^T\boldsymbol{V}) < 1$, holds for all choices of T, the only unique solution to the linear system is $(\boldsymbol{x}^*, \boldsymbol{u}^*)$ with the condition that any feasible $\boldsymbol{u}$ is s-sparse and any feasible $\boldsymbol{x}$ is in* $\text{Ker}(\boldsymbol{A})^\perp$,

Of interest is the identification of simple conditions such that $\sigma_1(\boldsymbol{U}^T\boldsymbol{V}) < 1$. The following theorem proposes one such condition to establish uniqueness of the dense and sparse coding problem.

Theorem 5 *Assume that there exists at least one solution to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, namely the pair $(\boldsymbol{x}^*, \boldsymbol{u}^*)$. Let S, with $|S| = s$, denote the support of $\boldsymbol{u}^*$. Assume that $\boldsymbol{B}_S$ has full column rank . Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times q}$ be matrices whose columns are the orthonormal bases of* $\text{Col}(\boldsymbol{A})$ *and* $\text{Col}(\boldsymbol{B}_S)$ *respectively. Let $\max_{i,j} |(\boldsymbol{U}^T\boldsymbol{V})_{i,j}| = \mu$. If $s < \frac{1}{\sqrt{r}\mu}$, the only unique solution to the linear system, with the condition that any feasible s-sparse vector $\boldsymbol{u}$ is supported on S and any feasible $\boldsymbol{x}$ is in* $\text{Ker}(\boldsymbol{A})^\perp$, *is $(\boldsymbol{x}^*, \boldsymbol{u}^*)$.*

Proof 3 *It suffices to show that $\sigma_1 < 1$. Noting that $\sigma_1 = ||\boldsymbol{U}^T\boldsymbol{V}||_2$, we use the following matrix norm inequality $||\boldsymbol{U}^T\boldsymbol{V}||_2 \leq \sqrt{r}||\boldsymbol{U}^T\boldsymbol{V}||_\infty$ as follows: $\sigma_1 \leq \sqrt{r}\,||\boldsymbol{U}^T\boldsymbol{V}||_\infty \leq \sqrt{r}\mu s < 1$.*

The constant μ is the coherence of the matrix $\boldsymbol{U}^T\boldsymbol{V}$ (Donoho et al., 2005; Tropp, 2004). The above result states that if the mutual coherence of $\boldsymbol{U}^T\boldsymbol{V}$ is small, we can accommodate increased sparsity of the underlying signal component $\boldsymbol{u}^*$. We note that, up to a scaling factor, $\sigma_1(\boldsymbol{U}^T\boldsymbol{V})$ is the block coherence of $\boldsymbol{U}$ and $\boldsymbol{V}$ (Eldar et al., 2010). However, unlike the condition in (Eldar et al., 2010), we don't restrict the dictionaries $\boldsymbol{A}$ and $\boldsymbol{B}$ to have linearly independent columns. In the next subsection, we propose a convex program to recover the dense and sparse vectors. Theorem 8 establishes uniqueness and complexity results for the proposed optimization program.

2.2 Dense and sparse recovery via convex optimization

Given that the dense and sparse coding problem seeks a dense vector $\boldsymbol{x}^*$ and a sparse solution $\boldsymbol{u}^*$, with measurements given as $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{B}\boldsymbol{u}^*$, we propose the following convex optimization program

$$\min_{\boldsymbol{x},\boldsymbol{u}} ||\boldsymbol{A}\boldsymbol{x}||_2^2 + ||\boldsymbol{u}||_1 \text{ s.t. } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}. \quad (2)$$

In this section, we show that, under certain conditions, the above minimization problem admits a unique solution. Our proof is a non-trivial adaptation of the existing analysis in (Kueng and Gross, 2014) for the anistropic compressive sensing problem. This analysis is based on a single measurement matrix and can not be directly applied to our scenario. Let $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ be a sequence of zero-mean i.i.d random vectors drawn from some distribution F on $\mathbb{R}^p$ and let $\boldsymbol{b}_1, ..., \boldsymbol{b}_m$ be a sequence of zero-mean i.i.d random vectors drawn from some distribution G on $\mathbb{R}^n$. We can eliminate the dense component in the linear constraint by projecting the vector $\boldsymbol{y}$ onto the orthogonal complement of $\text{Col}(\boldsymbol{A})$ to obtain $\mathcal{P}_{\text{Col}(\boldsymbol{A})^\perp}(\boldsymbol{y}) = \mathcal{P}_{\text{Col}(\boldsymbol{A})^\perp}(\boldsymbol{B}\boldsymbol{u})$. With this, the matrix $\mathcal{P}_{\text{Col}(\boldsymbol{A})^\perp}(\boldsymbol{B})$ is central in the analysis to follow. We define the matrix $\boldsymbol{C}$ $\boldsymbol{C} = \frac{1}{\sqrt{m}}\sum_{i=1}^m \boldsymbol{e}_i\boldsymbol{c}_i^T$ where $\boldsymbol{c}_i = [\mathcal{P}_{\text{Col}(A)^\perp}(\boldsymbol{B})]^T\boldsymbol{e}_i$ denotes the i-th measurement vector corresponding to a row of this matrix. Further technical discussion on the matrix $\boldsymbol{C}$ is deferred to **Appendix B**. We use the measurement matrix $\boldsymbol{C}$ introduced above and adapt the anisotropic compressive sensing theory in (Kueng and Gross, 2014) to analyze uniqueness of the proposed program. Below, we give brief background to this theory highlighting important assumptions and results following the notation closely therein.

Anisotropic compressive sensing: Given a sequence of zero-mean i.i.d random vectors $\boldsymbol{d}_1, ..., \boldsymbol{d}_m$ drawn from some distribution F on $\mathbb{R}^n$, with measurements $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{u}^*$, the anisotropic compressive sensing problem studies the following optimization program

$$\min_{\boldsymbol{u}} ||\boldsymbol{u}||_1 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{D}\boldsymbol{u}, \quad (3)$$

where $\boldsymbol{D} = \frac{1}{\sqrt{m}}\sum_{i=1}^m \boldsymbol{e}_i\boldsymbol{d}_i^T$ and $\boldsymbol{u}^*$ is the sought-out sparse solution. The analysis makes three important assumptions.

Completeness: The covariance matrix $\boldsymbol{\Sigma}$ is invertible with condition number denoted by κ.

Incoherence: The incoherence parameter is the smallest number ν such that

$$\max_{1\leq i\leq n} |\langle \boldsymbol{d}, \boldsymbol{e}_i\rangle|^2 \leq \nu \text{ and } \quad \max_{1\leq i\leq n} |\langle \boldsymbol{d}, E[\boldsymbol{c}\boldsymbol{c}^*]^{-1}\boldsymbol{e}_i|^2 \leq \nu \quad (4)$$

hold almost surely.

Conditioning of the covariance matrix: We start with the following definition of the s-sparse condition number restated from (Kueng and Gross, 2014).

Definition 6 *(Kueng and Gross, 2014) The largest and smallest s-sparse eigenvalue of a matrix $\boldsymbol{X}$ are given by*

$$\lambda_{\max}(s, \boldsymbol{X}) := \max_{\boldsymbol{v}, ||\boldsymbol{v}||_0\leq s} \frac{||\boldsymbol{X}\boldsymbol{v}||_2}{||\boldsymbol{v}||_2}$$

$$\lambda_{\min}(s, \boldsymbol{X}) := \min_{\boldsymbol{v}, ||\boldsymbol{v}||_0\leq s} \frac{||\boldsymbol{X}\boldsymbol{v}||_2}{||\boldsymbol{v}||_2}.$$

The s-sparse condition number of $\boldsymbol{X}$ *is* $cond(s, \boldsymbol{X}) = \frac{\lambda_{\max}(s, \boldsymbol{X})}{\lambda_{\min}(s, \boldsymbol{X})}$.

Given these assumptions, the main result in (Kueng and Gross, 2014) reads

Theorem 7 *(Kueng and Gross, 2014) With* $\kappa_s = \max\{cond(s, \boldsymbol{\Sigma}), cond(s, \boldsymbol{\Sigma}^{-1})\}$ *let* $\boldsymbol{u} \in \mathbb{C}^n$ *be an s-sparse vector and let* $\omega \geq 1$. *If the number of measurements fulfills* $m \geq C\kappa_s \nu \omega^2 s \log n$, *then the solution* $\boldsymbol{u}$ *of the convex program* (3) *is unique and equal to* $\boldsymbol{u}^*$ *with probability at least* $1 - e^{-\omega}$.

The proof of Theorem 7 is based on the dual certificate approach. The idea is to first propose a dual certificate vector $\boldsymbol{v}$ with sufficient conditions that ensure uniqueness of the minimization problem. It then remains to construct the dual certificate satisfying the conditions. We seek a similar result for the uniqueness of the convex program corresponding to the dense and sparse coding model. However, the standard analysis can not be directly applied since it only considers a single measurement matrix. This requires us to analyze the matrix $\boldsymbol{C}$ introduced earlier.

The anisotropic compressive sensing analysis in (Kueng and Gross, 2014) assumes the following conditions on the dual certificate $\boldsymbol{v}$

$$||\boldsymbol{v}_S - \operatorname{sgn}(\boldsymbol{u}_S^*)||_2 \leq \tfrac{1}{4} \quad \text{and} \quad ||\boldsymbol{v}_{S^\perp}||_\infty \leq \tfrac{1}{4}. \tag{5}$$

The following condition follows from the assumptions in Theorem 7

$$||\boldsymbol{\Delta}_S||_2 \leq 2||\boldsymbol{\Delta}_{S^\perp}||_2, \tag{6}$$

where $\boldsymbol{\Delta} \in \operatorname{Ker}(\boldsymbol{D})$. The conditions (5) and (6) will be used in the proof of our main result. The main part of the technical analysis in (Kueng and Gross, 2014) is using the assumptions in Theorem 7 and showing that the above conditions (5) and (6) hold with high probability.

Main result: Using the the background discussed above, we assume completeness, incoherence, and conditioning of the covariance matrix $\boldsymbol{\Sigma}$. Our main result is stated below.

Theorem 8 *Assume that there exists at least one solution to* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, *namely the pair* $(\boldsymbol{x}^*, \boldsymbol{u}^*)$. *Let* $\omega \geq 1$ *and define* $\kappa_s = \max\{cond(s, \boldsymbol{\Sigma}), cond(s, \boldsymbol{\Sigma}^{-1})\}$. *Assume the two conditions*

$$||\boldsymbol{B}_S^T \boldsymbol{A}|| \leq \frac{1}{32||\boldsymbol{x}^*||_2}, \quad ||\boldsymbol{B}_{S^\perp}^T \boldsymbol{A}||_\infty \leq \frac{1}{32||\boldsymbol{x}^*||_\infty}. \tag{7}$$

If the number of measurements fulfills $m \geq C\kappa_s \nu \omega^2 s \log n$, *then the solution of the convex program* (2) *is unique and equal to* $(\boldsymbol{x}^*, \boldsymbol{u}^*)$ *with probability at least* $1 - e^{-\omega}$.

Proof sketch 1 *Consider a feasible solution pair* $(\boldsymbol{x}^* + \boldsymbol{\delta_1}, \boldsymbol{u}^* + \boldsymbol{\delta_2})$ *and let the function* $f(\boldsymbol{x}, \boldsymbol{u})$ *denote the objective in the optimization program. The idea of the proof is to show that any feasible solution is not minimal in the objective value,* $f(\boldsymbol{x}^* + \boldsymbol{\delta_1}, \boldsymbol{u}^* + \boldsymbol{\delta_2}) > f(\boldsymbol{x}, \boldsymbol{u})$. *Using duality of the* ℓ_1 *norm and characterization of the subgradient* $\boldsymbol{\Lambda}$ *of the* ℓ_1 *norm, we first show that* $f(\boldsymbol{x}^* + \boldsymbol{\delta_1}, \boldsymbol{u}^* + \boldsymbol{\delta_2}) > f(\boldsymbol{x}^*, \boldsymbol{u}^*) + \langle \operatorname{sgn}(\boldsymbol{u}_S^*) + \boldsymbol{\Lambda} - \boldsymbol{v} - 2\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{\delta_2} \rangle$ *where* $\boldsymbol{v} \in \operatorname{Col}(\boldsymbol{C}^T)$, *with* $\boldsymbol{C} = \boldsymbol{B} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{B}$ *denoting the dual certificate. It then remains to show that the term* $\langle \operatorname{sgn}(\boldsymbol{u}_S^*) + \boldsymbol{\Lambda} - \boldsymbol{v} - 2\boldsymbol{B}^T\boldsymbol{A}\boldsymbol{x}^*, \boldsymbol{\delta_2} \rangle$ *is positive. To show this, we further analyze this term and make use of the assumptions of the theorem, the dual certificate conditions* (5), *and the deviation inequality in* (6) *to arrive at the desired result. For a complete proof, see* ***Appendix B***.

Complexity compared to ℓ_1 minimization: The sample complexity of solving the convex program corresponding to the dense and sparse coding problem is larger than that of ℓ_1 minimization for the compressive sensing problem. Essentially, the constants κ_s and ν in our analysis are expected to scale with $p + n$, in contrast to the compressive sensing analysis where they scale with n.

3 Experiments

3.1 Phase transition curves

We generate *phase transition curves* and present how the success rate of the recovery, using the proposed model, changes under different scenarios. To generate the data, we fix the number of columns of $\boldsymbol{B}$ to be $n = 100$. Then, we vary the sampling ratio $\sigma = \frac{m}{n+p} \in [0.05, 0.95]$ and the sparsity ratio $\rho = \frac{s}{m}$ in the same range. The sensing matrix in our model is $[\boldsymbol{A} \quad \boldsymbol{B}]$, hence the apparent difference in the definition of σ compared to "traditional" compressive sensing. In the case where we revert to the compressive sensing scenario ($p = 0$), the ratios coincide.

We generate random matrices $\boldsymbol{A} \in \mathbb{R}^{m \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ whose columns have expected unit norm. The vector $\boldsymbol{u} \in \mathbb{R}^n$ has s randomly chosen indices, whose entries are drawn according to a standard normal distribution, and $\boldsymbol{x} \in \mathbb{R}^p$ is generated as follows: we generate a random vector $\boldsymbol{\gamma} \in \mathbb{R}^m$, and then construct $\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{\gamma}$. The construction ensures that $\boldsymbol{x}$ does not belong in the null space of $\boldsymbol{A}$, and hence ignores trivial solutions with respect to this dense component. We normalize both $\boldsymbol{x}$ and $\boldsymbol{u}$ to have unit norm, and generate the measurement vector $\boldsymbol{y} \in \mathbb{R}^m$ as $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$. We solve the convex optimization problem in (2) to obtain the numerical solution pair $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}})$ using `CVXPY`, and

register a successful recovery if both $\frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \leq \epsilon$ and $\frac{\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \leq \epsilon$, with $\epsilon = 10^{-3}$. For each choice of σ and ρ we average 100 independent runs to estimate the success rate.

Figure 1 shows the phase transition curves for $p \in \{0.1m, 0.5m\}$ to highlight different ratios between p and n. We observe that increasing p leads to a deterioration in performance. This is expected, as this creates a greater *overlap* on the spaces spanned by $\boldsymbol{A}$ and $\boldsymbol{B}$. We can view our model as explicitly modeling the noise of the system. In such a case, the number of columns of $\boldsymbol{A}$ explicitly encodes the complexity of the noise model: as p increases, so does the span of the noise space.

Extending the signal processing interpretation, note that we model the noise signal $\boldsymbol{x}$ as a dense vector, which can be seen as encoding smooth areas of the signal that correspond to *low-frequency* components. On the contrary, the signal $\boldsymbol{u}$ has, by construction, a sparse structure, containing *high-frequency* information, an interpretation that will be further validated in the next subsection. Further numerical experiments comparing the dense and sparse coding model to noisy compressive sensing can be found in **Appendix C**.

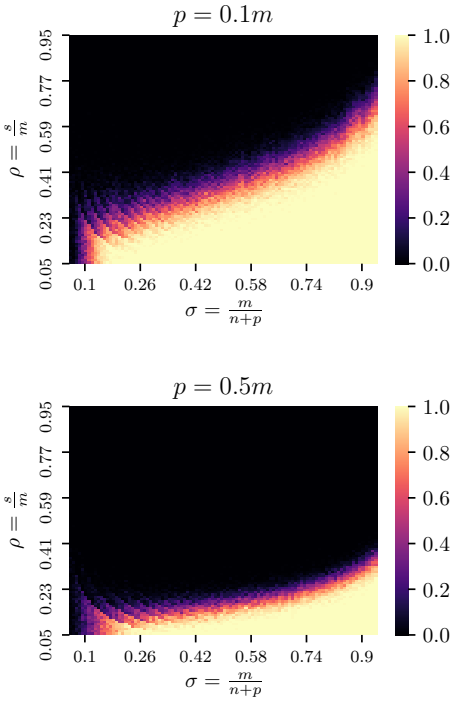

Figure 1: Phase transition curves for $p = 0.1m$ (top) and $p = 0.5m$ (bottom).

3.2 Classification and image denoising

We formulate the dense and sparse dictionary learning problem as minimizing the objective $\min_{\boldsymbol{A},\boldsymbol{B},\{\boldsymbol{x}^j\}_{j=1}^J,\{\boldsymbol{u}^j\}_{j=1}^J} \sum_{j=1}^J \frac{1}{2}\|\boldsymbol{y}^j - \boldsymbol{A}\boldsymbol{x}^j - \boldsymbol{B}\boldsymbol{u}^j\|_2^2 + \frac{1}{2\lambda_x}\|\boldsymbol{A}\boldsymbol{x}^j\|_2^2 + \lambda_u\|\boldsymbol{u}^j\|_1$, where J is the number of images, λ_x controls the smoothness of $\boldsymbol{A}\boldsymbol{x}^j$ and λ_u controls the degree of sparsity. Based on the objective, we use deep unfolding to construct a unfolding neural network (Tolooshams et al., 2020; Gregor and Lecun, 2010), which we term the dense and sparse autoencoder (DenSaE), tailored to learning the dictionaries from the dense and sparse model. The encoder maps $\boldsymbol{y}^j$ into a dense vector $\boldsymbol{x}_T^j$ and a sparse one $\boldsymbol{u}_T^j$ by unfolding T proximal gradient iterations. The decoder reconstructs the image. For classification, we use $\boldsymbol{u}_T$ and $\boldsymbol{x}_T$ as inputs to a linear classifier $\boldsymbol{C}$ that maps them to the predicted class $\hat{\boldsymbol{q}}$. We learn the dictionaries $\boldsymbol{A}$ and $\boldsymbol{B}$, as well as the classifier $\boldsymbol{C}$, by minimizing the weighted reconstruction (AE) and classification (Logistic) loss (i.e., $(1-\beta)$ AE + β Logistic). Figure 2 shows the DenSaE architecture (for details see **Appendix D**).

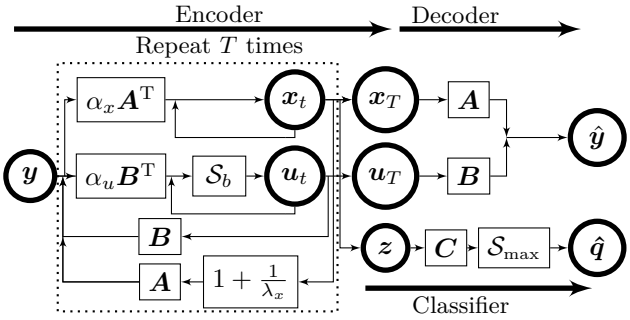

Figure 2: DenSaE. The vector $\boldsymbol{z}$ is normalized stacked features with a column of 1 (i.e., $\boldsymbol{z}^{\mathrm{T}} = [\boldsymbol{1}, \frac{[\boldsymbol{x}_T, \boldsymbol{u}_T]}{\|[\boldsymbol{x}_T, \boldsymbol{u}_T]\|}]$), $\mathcal{S}_b$ is soft-thresholding, and $\mathcal{S}_{\max}$ is softmax.

We examined the following questions

(a) How do the discriminative, reconstruction, and denoising capabilities change as we vary the number of filters in $\boldsymbol{A}$ vs. $\boldsymbol{B}$?
(b) What is the performance of DenSaE compared to sparse coding networks?
(c) What data characteristics does the model capture?

As baselines, we trained two variants, $\text{CSCNet}_{\text{hyp}}^{\text{tied}}$ and $\text{CSCNet}_{\text{LS}}^{\text{tied}}$, of CSCNet (Simon and Elad, 2019), an architecture tailored to dictionary learning for the sparse coding problem. In $\text{CSCNet}_{\text{hyp}}^{\text{tied}}$, the bias is a shared hyper-parameter. In $\text{CSCNet}_{\text{LS}}^{\text{tied}}$, we learn a different bias for each filter by minimizing the reconstruction loss. When the dictionaries are non-convolutional, we call the network SCNet.

Table 1: DenSaE's performance on MNIST test dataset from both disjoint (D) and joint (J_β) training.

		$SCNet^{LS}_{hyp}$	$SCNet^{tied}_{hyp}$	$\frac{5\boldsymbol{A}}{395\boldsymbol{B}}$	$\frac{25\boldsymbol{A}}{375\boldsymbol{B}}$	$\frac{200\boldsymbol{A}}{200\boldsymbol{B}}$
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ model	-	-	1.25	6.25	50
D	Acc.	94.16	98.32	98.18	98.18	96.98
	Rec.	1.95	6.80	6.83	6.30	3.04
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ class	-	-	0	0	0
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ rec.	-	-	8	28	58
$J_{0.75}$	Acc.	96.91	98.18	98.19	98.23	97.64
	Rec.	2.17	1.24	0.75	1.11	**0.51**
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ class	-	-	8	8	84
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ rec.	-	-	8	36	8
J_1	Acc.	96.06	98.59	**98.61**	98.56	98.40
	Rec.	71.20	47.70	32.61	30.20	25.57
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ class	-	-	16	46	42
	$\frac{\boldsymbol{A}}{\boldsymbol{A}+\boldsymbol{B}}$ rec.	-	-	0	2	4

3.2.1 DenSaE strikes a balance between discriminative capability and reconstruction

We study the case when DenSaE is trained on the MNIST dataset for joint reconstruction and classification purposes. We show a) how the explicit imposition of sparse and dense representations in DenSaE helps to balance discriminative and representation power, and b) that DenSaE outperforms SCNet. We warm start the training of the classifier using dictionaries obtained first training the autoencoder, i.e., with $\beta = 0$.

Characteristics of the representations $\boldsymbol{x}_T$ and $\boldsymbol{u}_T$: To evaluate the discriminative power of the representations learned by only training the autoencoder, we first trained the classifier *given* the representations (i.e., first train $\boldsymbol{A}$ and $\boldsymbol{B}$ with $\beta = 0$, then train $\boldsymbol{C}$ with $\beta = 1$). We call this disjoint training. The first four rows of section D from Table 1 show, respectively, the classification accuracy (Acc.), ℓ_2 reconstruction loss (Rec.), and the relative contributions, expressed as a percentage, of the dense or sparse representations to classification and reconstruction for disjoint training. Each col of $[\boldsymbol{A}\ \boldsymbol{B}]$, and of $\boldsymbol{C}$, corresponds to either a dense or a sparse feature. For reconstruction, we find the indices of the 50 most important columns and report the proportion of these that represent dense features. For each of the 10 classes (rows of $\boldsymbol{C}$), we find the indices of the 5 most important columns (features) and compute the proportion of the total of 50 indices that represent dense features. The first row of Table 1 shows the proportion of rows of $[\boldsymbol{A}\ \boldsymbol{B}]$ that represent dense features. Comparing this row, respectively to the third and fourth row of section D reveals the importance of $\boldsymbol{x}$ for reconstruction, and of $\boldsymbol{u}$ for classification. Indeed, the first two rows of section D show that, as the proportion of dense features increases, DenSaE gains reconstruction capability but results in a lower classification accuracy. Moreover, in DenSaE, the most important features in classification are all from $\boldsymbol{B}$, and the contribution of $\boldsymbol{A}$ in reconstruction is greater than its percentage in the model, which clearly demonstrates that dense and sparse coding autoencoders balance discriminative and representation power.

The table also shows that DenSaE outperforms $SCNet^{LS}_{hyp}$ in classification and $SCNet^{tied}_{hyp}$ in reconstruction. We observed that in the absence of noise, training $SCNet^{LS}_{hyp}$ results in dense features with negative biases, hence, making its performance close to DenSaE with large number of atoms in $\boldsymbol{A}$. We see that $SCNet^{LS}_{hyp}$ in absence of a supervised classification loss fails to learn discriminative features useful for classification. On the other hand, enforcing sparsity in $SCNet^{tied}_{hyp}$ suggests that sparse representations are useful for classification.

How do roles of $\boldsymbol{x}_T$ and $\boldsymbol{u}_T$ change as we vary β in joint training?: In joint training of the autoencoder and the classifier, it is natural to expect that the reconstruction loss should increase compared to disjoint training. This is indeed the case for $SCNet^{LS}_{hyp}$; as we go from disjoint to joint training and as β increases (Table 1, sections labeled J), the reconstruction loss increases and classification accuracy has an overall increase. However, for $\beta < 1$, joint training of both networks that enforce some sparsity on their representations, $SCNet^{tied}_{hyp}$ and DenSaE, improves reconstruction and classification. Moreover, as we increase the importance of classification loss (i.e., increase β), the contribution of dense representations decreases in reconstruction and increases in discrimination.

For purely discriminative training ($\beta = 1$), DenSaE outperforms both $SCNet^{LS}_{hyp}$ and $SCNet^{tied}_{hyp}$ in classification accuracy and representation capability. We speculate that this likely results from the fact that, by construction the encoder from DenSaE seeks to produce two sets of representations: namely a dense one, mostly important for reconstruction and a sparse one, useful for classification. In some sense, the dense component acts as a prior that promotes good reconstruction. More detailed results can be found in **AppendixD**.

Remarks: As our network is non-convolutional, we do not compare it to the state-of-the-art, a convolutional network. We do not compare our results with the network in (Rolfe and LeCun, 2013) as that work does not report reconstruction loss and it involves a sparsity enforcing loss that change the learning behaviour.

3.2.2 Denoising

We trained DenSaE for supervised image denoising when $\beta = 0$ using BSD432 and tested it on BSD68 (Martin et al., 2001) (see **Appendix D** for details). We

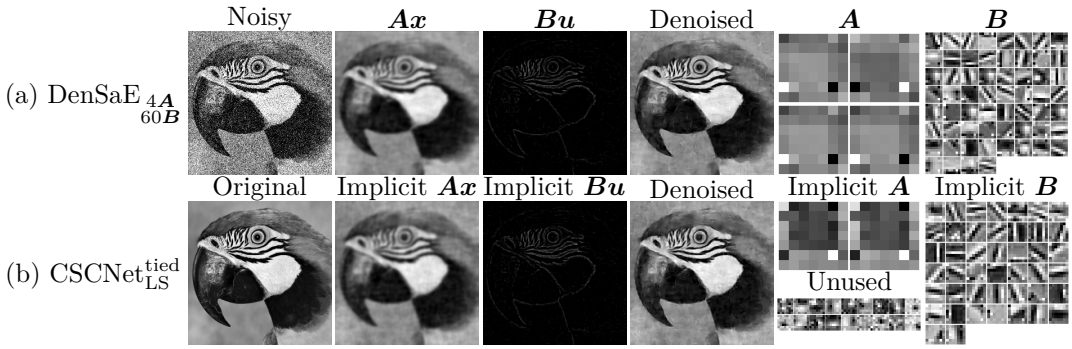

Figure 3: Visualization of a test image for $\tau = 50$. (a) DenSaE $(4\boldsymbol{A}, 60\boldsymbol{B})$, (b) CSCNet$_{\text{LS}}^{\text{tied}}$.

Table 2: DenSaE's denoising performance on test BSD68 as the ratio of filters in $\boldsymbol{A}$ and $\boldsymbol{B}$ changes.

τ	$1\boldsymbol{A}63\boldsymbol{B}$	$4\boldsymbol{A}60\boldsymbol{B}$	$8\boldsymbol{A}56\boldsymbol{B}$	$16\boldsymbol{A}48\boldsymbol{B}$	$32\boldsymbol{A}32\boldsymbol{B}$
15	**30.21**	30.18	30.18	30.14	29.89
25	**27.70**	**27.70**	27.65	27.56	27.26
50	**24.81**	**24.81**	24.43	24.44	23.68
75	23.31	**23.33**	23.09	22.09	20.09

Table 3: DenSaE vs. CSCNet on test BSD68.

τ	DenSaE	CSCNet$_{\text{hyp}}^{\text{tied}}$	CSCNet$_{\text{LS}}^{\text{tied}}$
15	30.21	30.12	**30.34**
25	27.70	27.51	**27.75**
50	**24.81**	24.54	**24.81**
75	**23.33**	22.83	23.32

varied the ratio of number of filters in $\boldsymbol{A}$ and $\boldsymbol{B}$ as the overall number of filters was kept constant. We evaluate the model in the presence of Gaussian noise with standard deviation of $\tau = \{15, 25, 50, 75\}$.

Ratio of number of filters in $\boldsymbol{A}$ and $\boldsymbol{B}$: Table 2 shows that the smaller the number of filters associated with $\boldsymbol{A}$, the better DenSaE can denoise images. We hypothesize that this is a direct consequence of our findings from **Section 2** that the smaller the number of columns of $\boldsymbol{A}$, the easier the recovery $\boldsymbol{x}$ and $\boldsymbol{u}$.

Dense and sparse coding vs. sparse coding: Table 3 shows that DenSaE (best network from Table 2) denoises images better than CSCNet$_{\text{hyp}}^{\text{tied}}$, suggesting that the dense and sparse coding model represents images better than sparse coding.

Dictionary characteristics: Figure 3(a) shows the decomposition of a noisy test image ($\tau = 50$) by DenSaE. The figure demonstrates that $\boldsymbol{Ax}$ captures low-frequency content while $\boldsymbol{Bu}$ captures high-frequency details (edges). This is corroborated by the smoothness of the filters associated with $\boldsymbol{A}$, and the Gabor-like nature of those associated with $\boldsymbol{B}$ (Mehrotra et al., 1992). We observed similar performance when we tuned λ_x, and found that, as λ_x decreases, $\boldsymbol{Ax}$ captures a lower frequencies, and $\boldsymbol{Bu}$ a broader range.

CSCNet implicitly learns $\boldsymbol{Ax}+\boldsymbol{Bu}$ model in the presence of noise: We observed that CSCNet$_{\text{LS}}^{\text{tied}}$ comprises three groups of filters: one with small bias, one with intermediate ones, and a third with large values (see **Appendix D** for bias visualizations). We found that the feature maps associated with the large bias values are all zero. Moreover, the majority of features are associated with intermediate bias values, and are sparse, in contrast to the small number of feature maps with small bias values, which are dense. These observations suggest that *autoencoder implementing the sparse coding model ($\boldsymbol{y} = \boldsymbol{Bu}$), when learning the biases by minimizing reconstruction error, implicitly perform two functions.* First, they select the optimal number of filters. Second, they partition the filters into two groups: one that yields a dense representation of the input, and another that yields a sparse one. In other words, the architectures trained in this manner *implicitly learn the dense and sparse coding model* ($\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{Bu}$). Figure 3(b) shows the filters.

4 Conclusions

This paper proposed a novel dense and sparse coding model for a flexible representation of a signal as $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{Bu}$. Our first result gives a verifiable condition that guarantees uniqueness of the model. Our second result uses tools from RIPless compressed sensing to show that, with sufficiently many linear measurements, a convex program with ℓ_1 and ℓ_2 regularizations can recover the components $\boldsymbol{x}$ and $\boldsymbol{u}$ uniquely with high probability. Numerical experiments on synthetic data confirm our observations.

We proposed a dense and sparse autoencoder, DenSaE, tailored to dictionary learning for the $\boldsymbol{Ax} + \boldsymbol{Bu}$ model. DenSaE, naturally decomposing signals into low- and high-frequency components, provides a balance between learning dense representations that are useful for reconstruction and discriminative sparse representations. We showed the superiority of DenSaE to sparse autoencoders for data reconstruction and its competitive performance in classification.

References

M. Aharon, M. Elad, and A. Bruckstein, "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–22, 2006.

J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 791–804, 2011.

B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

C. Garcia-Cardona and B. Wohlberg, "Convolutional dictionary learning: A comparative review and new algorithms," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366–81, Sep. 2018.

K. Gregor and Y. Lecun, "Learning fast approximations of sparse coding," in *Proc. International Conference on Machine Learning (ICML)*, 2010, pp. 399–406.

J. T. Rolfe and Y. LeCun, "Discriminative recurrent sparse auto-encoders," *arXiv preprint arXiv:1301.3775*, 2013.

D. Simon and M. Elad, "Rethinking the csc model for natural images," in *Proc. Advances in Neural Information Processing Systems*, 2019, pp. 2271–2281.

B. Tolooshams, S. Dey, and D. Ba, "Deep residual autoencoders for expectation maximization-inspired dictionary learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2020.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, vol. 11, no. 12, 2010.

B. Epstein, R. Meir, and T. Michaeli, "Joint autoencoders: a flexible meta-learning framework," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 494–509.

B. Epstein and R. Meir, "Generalization bounds for unsupervised and semi-supervised learning with autoencoders," *arXiv preprint arXiv:1902.01449*, 2019.

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*, pp. 1–1, 2019.

J. Zazo, B. Tolooshams, and D. Ba, "Convolutional dictionary learning in hierarchical networks," in *Proc. 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019, pp. 131–135.

D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE transactions on information theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2558–2567, 2002.

D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

M. Soltani and C. Hegde, "Fast algorithms for demixing sparse signals from nonlinear observations," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4209–4222, 2017.

C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei, "Recovery of sparsely corrupted signals," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3115–3130, 2011.

C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," *Applied and Computational Harmonic Analysis*, vol. 37, no. 1, pp. 12–35, 2014.

E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

L. Lian, A. Liu, and V. K. Lau, "Weighted lasso for sparse recovery with statistical prior support information," *IEEE Transactions on Signal Processing*, vol. 66, no. 6, pp. 1607–1618, 2018.

H. Mansour and R. Saab, "Recovery analysis for weighted l1-minimization using the null space property," *Applied and Computational Harmonic Analysis*, vol. 43, no. 1, pp. 23–38, 2017.

M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)," *Applied and computational harmonic analysis*, vol. 19, no. 3, pp. 340–358, 2005.

A. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of computation*, vol. 27, no. 123, pp. 579–594, 1973.

D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on information theory*, vol. 52, no. 1, pp. 6–18, 2005.

J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

R. Kueng and D. Gross, "Ripless compressed sensing from anisotropic measurements," *Linear Algebra and its Applications*, vol. 441, pp. 110–123, 2014.

D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

R. Mehrotra, K. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern Recognition*, vol. 25, no. 12, pp. 1479–94, 1992.