Digital Signal Processing ••• (••••) •••–•••

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Deep fully-connected networks for video compressive sensing

Michael Iliadis[†] *,[1], Leonidas Spinoulas[†], Aggelos K. Katsaggelos

Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208-3118, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Video compressive sensing
Deep neural networks
Fully-connected networks

ABSTRACT

In this work we present a deep learning framework for video compressive sensing. The proposed formulation enables recovery of video frames in a few seconds at significantly improved reconstruction quality compared to previous approaches. Our investigation starts by learning a linear mapping between video sequences and corresponding measured frames which turns out to provide promising results. We then extend the linear formulation to deep fully-connected networks and explore the performance gains using deeper architectures. Our analysis is always driven by the applicability of the proposed framework on existing compressive video architectures. Extensive simulations on several video sequences document the superiority of our approach both quantitatively and qualitatively. Finally, our analysis offers insights into understanding how dataset sizes and number of layers affect reconstruction performance while raising a few points for future investigation.

1. Introduction

The subdivision of time by motion picture cameras, the frame rate, limits the temporal resolution of a camera system. Even though frame rate increase above 30 Hz may be imperceptible to human eyes, high speed motion picture capture has long been a goal in scientific imaging and cinematography communities. Despite the increasing availability of high speed cameras through the reduction of hardware prices, fundamental restrictions still limit the maximum achievable frame rates.

Video compressive sensing (CS) aims at increasing the temporal resolution of a sensor by incorporating additional hardware components to the camera architecture and employing powerful computational techniques for high speed video reconstruction. The additional components operate at higher frame rates than the camera's native temporal resolution giving rise to low frame rate multiplexed measurements which can later be decoded to extract the unknown observed high speed video sequence. Despite its use for high speed motion capture [1], video CS also has applications to coherent imaging (e.g., holography) for tracking high-speed events [2] (e.g., particle tracking, observing moving biological samples). The benefits of video CS are even more pronounced for non-visible light applications where high speed cameras are rarely available or prohibitively expensive (e.g., millimeter-wave imaging, infrared imaging) [3,4].

Video CS comes in two incarnations, namely, spatial CS and temporal CS. Spatial video CS architectures stem from the well-known single-pixel-camera [5], which performs spatial multiplexing per measurement, and enable video recovery by expediting the capturing process. They either employ fast readout circuitry to capture information at video rates [6] or parallelize the single-pixel architecture using multiple sensors, each one responsible for sampling a separate spatial area of the scene [4,7].

In this work, we focus on temporal CS where multiplexing occurs across the time dimension. Fig. 1 depicts this process, where a spatio-temporal volume of size $W_f \times H_f \times t = N_f$ is modulated by t binary random masks during the exposure time of a single capture, giving rise to a coded frame of size $W_f \times H_f = M_f$.

We denote the vectorized versions of the unknown signal and the captured frame as $\mathbf{x} : N_f \times 1$ and $\mathbf{y} : M_f \times 1$, respectively. Each vectorized sampling mask is expressed as $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_t$ giving rise to the measurement model

$$\mathbf{y} = \Phi \mathbf{x}, \tag{1}$$

where $\Phi = [diag(\boldsymbol{\phi}_1), \ldots, diag(\boldsymbol{\phi}_t)] : M_f \times N_f$ and $diag(\cdot)$ creates a diagonal matrix from its vector argument.

Various successful temporal CS architectures have been proposed. Their differences mainly involve the implementation of the random masks on the optical path (i.e., the measurement matrix in Fig. 1). Digital micromirror devices (DMD), spatial light modulators (SLM) and liquid crystal on silicon (LCoS) were used in [4, 7–10] while translating printed masks were employed in [11,12]. Moreover, a few architectures have eliminated additional optical elements by directly programming the chip's readout mode through hardware circuitry modifications [13–15].

* Corresponding author.
E-mail address: miliad@u.northwestern.edu (M. Iliadis).
[1] The † next to the author names denotes equal contribution.

https://doi.org/10.1016/j.dsp.2017.09.010
1051-2004/

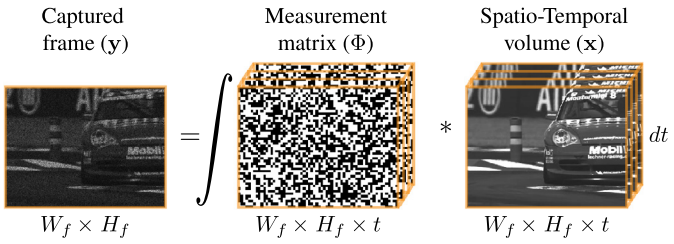

Fig. 1. Temporal compressive sensing measurement model.

Despite their reasonable performance, temporal CS architectures lack practicality. The main drawback is that existing reconstruction algorithms (e.g., using sparsity models [4,16], combining sparsity and dictionary learning [9] or using Gaussian mixture models [17, 18]) are often too computationally intensive, rendering the reconstruction process painfully slow. Even with parallel processing, recovery times make video CS prohibitive for modern commercial camera architectures.

In this work, we address this problem by employing deep learning and show that video frames can be recovered in a few seconds at significantly improved reconstruction quality compared to existing approaches.

Our contributions are summarized as follows:

1. We present the first deep learning architecture for temporal video CS reconstruction approach, based on fully-connected neural networks, which learns to map directly temporal CS measurements to video frames. For such task to be practical, a measurement mask with a repeated pattern is proposed.
2. We show that a simple linear regression-based approach learns to reconstruct video frames adequately at a minimal computational cost. Such reconstruction could be used as an initial point to other video CS algorithms.
3. The learning parading is extended to deeper architectures exhibiting reconstruction quality and computational cost improvements compared to previous methods.

2. Motivation and related work

Deep learning [19] is a burgeoning research field which has demonstrated state-of-the-art performance in a multitude of machine learning and computer vision tasks, such as image recognition [20] or object detection [21].

In simple words, deep learning tries to mimic the human brain by training large multi-layer neural networks with vast amounts of training samples, describing a given task. Such networks have proven very successful in problems where analytical modeling is not easy or straightforward (e.g., a variety of computer vision tasks [22,23]).

The popularity of neural networks in recent years has led researchers to explore the capabilities of deep architectures even in problems where analytical models often exist and are well understood (e.g., restoration problems [24–26]). Even though performance improvement is not as pronounced as in classification problems, many proposed architectures have achieved state-of-the-art performance in problems such as deconvolution, denoising, inpainting, and super-resolution.

More specifically, investigators have employed a variety of architectures: deep fully-connected networks or multi-layer perceptrons (MLPs) [24,25]; stacked denoising auto-encoders (SDAEs) [26–29], which are MLPs whose layers are pre-trained to provide improved weight initialization; convolutional neural networks (CNNs) [7,30–34] and recurrent neural networks (RNNs) [35].

Based on such success in restoration problems, we wanted to explore the capabilities of deep learning for the video CS problem.

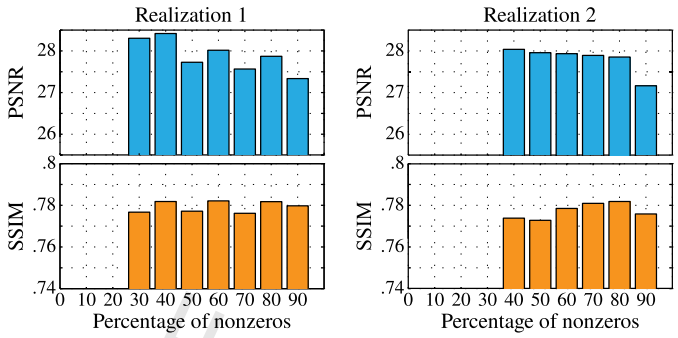

Fig. 2. Average reconstruction performance of linear mapping for 14 videos (unrelated to the training data), using measurement matrices Φ_p with varying percentages of nonzero elements.

However, the majority of existing architectures involve outputs whose dimensionality is smaller than the input (e.g., classification) or have the same size (e.g., denoising/deblurring). Hence, devising an architecture that estimates N_f unknowns, given M_f inputs, where $M_f \ll N_f$ is not necessarily straightforward.

Two recent studies, utilizing SDAEs [36] or CNNs [37], have been presented on spatial CS for still images exhibiting promising performance. Our work constitutes the first attempt to apply deep learning on temporal video CS. Our approach differs from prior 2D image restoration architectures [24,25] since we are recovering a 3D volume from 2D measurements.

3. Deep networks for compressed video

3.1. Linear mapping

We started our investigation by posing the question: can training data be used to find a linear mapping W such that $\mathbf{x} = W\mathbf{y}$? Essentially, this question asks for the inverse of Φ in equation (1) which, of course, does not exist. Clearly, such a matrix would be huge to store but, instead, one can apply the same logic on video blocks [9].

We collect a set of training video blocks denoted by $\mathbf{x}_i$, $i \in \mathbb{N}$ of size $w_p \times h_p \times t = N_p$. Therefore, the measurement model per block is now $\mathbf{y}_i = \Phi_p \mathbf{x}_i$ with size $M_p \times 1$, where $M_p = w_p \times h_p$ and Φ_p refers to the corresponding measurement matrix per block.

Collecting a set of N video blocks, we obtain the matrix equation

$$Y = \Phi_p X, \tag{2}$$

where $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and Φ_p is the same for all blocks. The linear mapping $X = W_p Y$ we are after can be calculated as

$$\min_{W_p} \|X - W_p Y\|_2^2 \rightarrow W_p = \left(XY^T\right)\left(YY^T\right)^{-1}, \tag{3}$$

where W_p is of size $N_p \times M_p$.

Intuitively, such an approach would not necessarily be expected to even provide a solution due to ill-posedness. However, it turns out that, if N is sufficiently large and the matrix Φ_p has at least one nonzero in each row (i.e., sampling each spatial location at least once over time), the estimation of $\mathbf{x}_i$'s by the $\mathbf{y}_i$'s provides surprisingly good performance.

Specifically, we obtain measurements from a test video sequence applying the same Φ_p per video block and then reconstruct all blocks using the learnt W_p. Fig. 2 depicts the average peak signal-to-noise ratio (PSNR) and structural similarity metric (SSIM) [38] for the reconstruction of 14 video sequences using 2 different realizations of the random binary matrix Φ_p for varying percentages of nonzero elements. The empty bars for 10–20% and

Please cite this article in press as: M. Iliadis et al., Deep fully-connected networks for video compressive sensing, Digit. Signal Process. (2017), https://doi.org/10.1016/j.dsp.2017.09.010