Neural Networks 29–30 (2012) 1–7

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Neural networks letter

Capturing significant events with neural networks

Harold Szu [a,*], Charles Hsu [b], Jeffrey Jenkins [a], Jefferson Willey [c], Joseph Landa [d]

[a] US Army NVESD, Fort Belvoir, VA, United States
[b] Trident Systems Inc., Fairfax, VA, United States
[c] US Naval Research Lab, Washington, DC, United States
[d] Briartek Inc., Alexandria, VA, United States

ARTICLE INFO

Article history:
Received 29 August 2011
Received in revised form 23 December 2011
Accepted 18 January 2012

Keywords:
Compressive sensing
Associative memory
Human visual system
Compressive video sampling

ABSTRACT

Smartphone video capture and transmission to the Web contributes to data pollution. In contrast, mammalian eyes sense all, capture only significant events, allowing us vividly recall the causalities. Likewise in our videos, we wish to skip redundancies and keep only significantly differences, as determined by real-time local medium filters. We construct a Picture Index (PI) of one's (center of gravity changes) among zeros (no changes) as *Motion Organized Sparseness (MOS)*. Only *non-overlapping time-ordered PI pair is admitted* in the outer-product *Associative* Memory (*AM*). Another outer product between PI and its image builds *Hetero-AM* (HAM) for fault tolerant retrievals.

1. Introduction

We wish to describe a spatiotemporal information sampling strategy for a small field of view handheld Smartphone. We begin with a common sense approach about video frame rates: "How many views or frames does a monkey need in order to tell a good zookeeper from a bad one?" Monkeys select 3 distinctive views, which we refer to as *m frames*: frontal, side and a 45° view (Giese & Poggio, 2003). Interestingly, humans need only $m = 2$ views when constructing a 3-D building from architectural blueprints, or for visualizing a human head. These kind of questions, posed by Giese and Poggio (2003), can be related to an important medical imaging application. The Compressive Sensing (CS) strategy in medical imaging may save the patient from un-wanted radiation exposure with a smaller number of m views of even smaller number of exposed pixels, which is counted as the l_p-norm $|\vec{x}|_p \equiv (\sum_{n=1}^{N} |x_n|^p)^{\frac{1}{p}}$, $0 \le p < 2$, where $p = 2$ is Least Mean Squared (LMS); $p = 1$ is Manhattan, or city-block, window-rim distance; and $p = 0$ counts the non-zero elements $\sum_{n=1}^{N} |\vec{X}_n|^0 = k$ (anything non-zero raised to zero power is 1 mathematically). The k-degree of sparseness is the sensing degree of freedom satisfying $\frac{k}{N} \ll 1$.

A million-dollar question remains in the medical imaging technology: how to block the imaging radiation for m optimal linear sparse combinations. We do not know a high precision technology for controlling the X-ray and Gallium contrast agents (cf. Concluding Remarks). Instead, we address the Smartphone data pollution challenge by exploiting an Artificial Neural Network (ANN) and Associative Memory (AM) learning technology. We wish to capture significantly meaningful m frames to render a video Cliff Note, without the usual post-processing time and labor costs.

We follow a mammalian vision, from Darwinian viewpoints, paying attention to significant and abrupt changes for mating, food, and survival reasons. For example, when driving at night, we watch for pedestrians, in the rain not random raindrops. When the visual stimuli received by both eyes agree at a moment in time, it is certainly a signal; if disagree, it could be noise and rejected. Such an experience of selective image fusion is effortless unsupervised learning to separate signal from noise. Combining the power of two eyes with the brain associative memory we can also solve sophisticated image sources de-mixing problems (cf. Concluding Remarks) (Szu & Kopriva, 2002).

Spotting Face App. Smartphone took 3 pictures of the same person of variable poses $\{\vec{x}_t,\ t = 1, 2, 3\}$ and 2 pictures of another person $\{\vec{y}_t,\ t = 1, 2\}$, etc., created a private FaceBook Web database: $[A] = [\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{y}_1, \vec{y}_2, \ldots]$. After a phone call meeting a friend in a football stadium, one may wish to turn on the spotting face app. The phone camera can match any incoming picture $\vec{Y}$ with the Smartphone database $[A]$, which is mathematically equivalent to an over-determined inverse, by fitting a highly redundant database $[A]$ with a sparse representation $\vec{X}$ of the incoming $\vec{Y}$ (identifying $\vec{Y}$ with known facial poses $[A]$ must be sparse $\vec{X}$ to be potentially unique).

$$\vec{Y}_N = [A]_{N,m}\vec{X}_m; \Longrightarrow \vec{y}_m = [B]_{m,m}\vec{X}_m; \tag{1a}$$

* Corresponding author. Tel.: +1 703 704 0532.
E-mail addresses: szuharoldh@gmail.com, harold.h.szu.civ@mail.mil (H. Szu).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.01.003

Use is made of a purely random sparse sensing matrix $[\phi]_{m,N}$ in Compressive Sensing

$$[B]_{m,m} \equiv [\phi]_{m,N}[A]_{N,m}; \quad \text{and} \quad \overrightarrow{y}_m \equiv [\phi]_{m,N}\overrightarrow{Y}_N. \tag{1b}$$

The sparse sensing matrix will be replaced by a video motion organized sparse sampling matrix $[\phi_s]$ in our video Compressive Sampling.

Use is made of a purely random sparse sensing matrix $[\phi]_{m,N}$ in Compressive Sensing *Single frame app*. Given an image acquisition rectangular matrix of m rows *and* N *columns* $[A]_{m,N}$ consisting of few known ones (transparent ones for keeping the pixels) per row among a dense sea of zeros (opaque zeros for rejecting pixels). Statistically speaking, the ones of each row will not be overlapped with the ones in the other row which is sparse and thus orthogonal in the statistical sense. Thus, the same linear algebra equation (1a) can be interpreted differently in single frame app. The column vector $\overrightarrow{Y}$ has m measured summary values (similarly to Monkey's need $m = 3$ views and human $m = 2$ views) of the unknown image vector $\overrightarrow{X}$ of N pixels. Given the input measurement vector $\overrightarrow{Y}$ and known sparse orthogonal acquisition matrix $[A]_{m,N}$, we must determine the unknown image $\overrightarrow{X}$ of N pixels. The question is what to do when there are $N - m$ missing conditions. In other words, finding $\overrightarrow{X}$ from $\overrightarrow{Y}$ becomes an ill-posed inverse problem.

Solving the ill-posed inverse requires a performance measure, e.g. LMS similarity min.$|\overrightarrow{X} - \overrightarrow{Y}|^2$ l_2-norm, together with a constraint at the minimum or sparse city-block distance: min.$|\overrightarrow{X}|$ of the l_1 norm, rather l_0, for computational tractability reasons. Without the constraint, the LMS is blind to all possible direction cosines within the hyper-sphere surface, called Penrose's pseudo-inverse: right-multiplier $[A]^{-1} \cong [A]^T([A][A]^T)^{-1}$; or left-multiplier $[A]^{-1} \cong ([A]^T[A])^{-1}[A]^T$. Indeed, using the sparseness constraint, solving the l_1-constrained l_2-optimization becomes a linear programming CS problem, as published by Emmanuel Candes of Caltech, Justin Romberg of GIT Technology, and Fields prize winner Terrence Tao of UCLA (Candes, Romberg, & Tao, 2006; Candes & Tao, 2006) as well as David Donoho of Stanford who adopted the pre-processing of wavelet sub-band codec before CS (Donoho, 2006). The Sparse Measurement Theorem stated that the sampling operator $[\Phi]$ has the Restricted Isometry Property if its matrix representation has k ones randomly distributed among zeros is bounded within: $\|[\Phi]\overrightarrow{X}\|/\|\overrightarrow{X}\| \cong O(1 \mp \delta_k)$; $m \cong 1.3k \ll N$. Instead of seeking a coarse inverse solution $\overrightarrow{X}$ located on the hyper-sphere LMS surface, CRTD sought after a sharper solution at a corner of the hypercube inscribed within the hyper-sphere surface, by imposing the sparseness constraint as a min.l_1-norm (by the true sparseness constraint at min.l_0-norm having N biaxial 2^N combinatorial choices).

Now we introduce our paper. There is more than one way to achieve the sparseness, a purely randomly way or an organized way. We choose the latter to assign the information meaning to the location of ones. These ones are selected sparsely by the significant changes of local Center of Gravity among neighborhood frames, and, otherwise, zeros. Furthermore, this sparse one among zeros is taken as the Picture Index $\overrightarrow{PI}_\tau$ representing the full resolution image $\overrightarrow{X}_\tau$ in the Massive Distributive Parallel (MDP) Hetero-Associative Memory:

$$[HAM] = \overrightarrow{X}_\tau \overrightarrow{PI}^T_\tau \tag{2}$$

where the superscript T denotes a transpose of a column vector to a row vector. The sampling rate can be adaptively decided by the C.G. changes of local scenery movement, from $\frac{1}{\triangle t} = 30$ Hz toward a few Hz or less, until the next $\overrightarrow{PI}_{\tau+\triangle t_\tau}$ is found and satisfied the orthogonality condition.

$$\langle \overrightarrow{PI}^T_\tau \overrightarrow{PI}_{\tau'} \rangle = \delta_{\tau,\tau'} \tag{3}$$

where the information flow $\tau' = t + \triangle t_t$; $t = \tau$, help selected and kept 2 frames form several by-passed frames. Thereby, we record this fact in the jump-over sequential storage index Associative Memory

$$[AM] = \overrightarrow{PI}_{\tau+\triangle t_\tau} \overrightarrow{PI}^T_\tau. \tag{4}$$

Since the input of a Picture Index to the index associative memory $[AM]$ can reproduce the next new picture index in a Fault Tolerant fashion, as the originally stored time-order.

$$\sigma_o([AM]\overrightarrow{PI}_\tau) = \overrightarrow{PI}_{\tau+\triangle t_\tau}, \tag{5}$$

where the well-known McCulloch–Pitts neuronal sigmoid logic is the neuronal two-state normalization (firing or not) that has the Boltzmann canonical ensemble form in terms of the Boltzmann constant K_B and brain Kelvin local temperature T:

$$y = \sigma_T(x-\theta) = \frac{\exp\left(\frac{x-\theta}{2K_BT}\right)}{\left[\exp\left(\frac{x-\theta}{2K_BT}\right) + \exp\left(-\frac{x-\theta}{2K_BT}\right)\right]} = \frac{1}{\left[1 + \exp\left(-\frac{x-\theta}{K_BT}\right)\right]}.$$

In a cool down local limit, $K_BT \Longrightarrow 0$, the sigmoid logic is reducible to Von Neumann binary logic: $1 \geqq \sigma_o(x-\theta) \geqq 0$, that could be more zeros than ones as the sparse representation. Moreover, a sequentially updated Hetero-AM storage is defined

$$[HAM] = [HAM] + \overrightarrow{X}_{\tau+\triangle t_\tau} \overrightarrow{PI}^T_{\tau+\triangle t_\tau}. \tag{6}$$

$[HAM]$ can be used to recover a high resolution $\overrightarrow{X}_{\tau+\triangle t_\tau}$ image at some equilibrium temperature T

$$\overrightarrow{X}_{\tau+\triangle t_\tau} = \sigma_T([HAM]\overrightarrow{PI}_{\tau+\triangle t_\tau}). \tag{7}$$

This strategy emulates the Hippocampus *AM* storage in the center of the brain. The *AM* and sigmoid logic is familiar to the neural network community, but its relationship to the current Compressive Sensing has not been elucidated before. In this adaptive or learning aspect, our approach has generalized the statistically purely random sparse pseudo-orthogonality. Our orthogonality is deterministically achieved by non-overlapping ones over zeros.

In Section 2, we will review the *AM* storage in terms of sparse matrix algebra of outer-product 'write' operations and the inner-product 'read' operations. The *Picture Index (PI)* is automatically produced by a video frame generated *Motion Organized Sparseness (MOS) which is* crucial for achieving non-overlapping orthogonality, and therefore the *fault tolerance* (*FT*) and generalization. Our approach is intended to be frame-selective and information-compressive sampling. We wish to eventually produce an automated '*Cliff Notes*' (not shown), which would merge all distinctive frames into a single (large) frame story line to aid human analysts, who otherwise have to manually sift through terabytes of data.